Chapter 6

Random Variables

case study

A Jury of Your Peers?

Are accused criminals in the United States entitled to a “jury of their peers”? Sort of. The Sixth Amendment to the U.S. Constitution begins, “In all criminal prosecutions, the accused shall enjoy the right to a speedy and public trial, by an impartial jury of the State and district wherein the crime shall have been committed. . . .” There is no mention of a “jury of your peers” in the Constitution or any of its amendments. However, an 1879 U.S. Supreme Court decision said that a jury should be chosen from a group “composed of the peers or equals [of the accused]; that is, of his neighbors, fellows, associates, persons having the same legal status in society as he holds.”[1]

To meet the Sixth Amendment requirement of impartiality, most courts start by randomly selecting a large jury pool from the citizens who live in the court’s jurisdiction. The jurors for a given trial are then chosen from the jury pool in a process known as voir dire. Each prospective juror answers a set of questions posed by the judge and the lawyers for both the prosecution and the defense. Depending on their answers, prospective jurors are excluded or seated on the jury.

In one case that made it all the way to the Supreme Court, a defense lawyer in Michigan challenged the process of selecting the jury pool in the trial of his accused client. Here are the facts:[2]

- About 7.28% of the citizens in the court’s jurisdiction were black.
- The jury pool had between 60 and 100 members, only 3 of whom were black.

Is it plausible that a jury pool with so few black citizens could be chosen just by chance?

By the end of this chapter, you will be ready to analyze the results of the jury selection process and to render your own verdict about the defense attorney’s challenge.

Introduction

Do you drink bottled water or tap water? According to a recent report in *U.S. Mayor Newspaper*, about 75% of people drink bottled water regularly. Some people do so because they believe bottled water is safer than tap water. (There's little evidence to support this belief.) Others say they prefer the taste of bottled water. Can people really tell the difference?

ACTIVITY | Bottled Water versus Tap Water

MATERIALS:
3 small paper cups per student; enough tap water for 2 cups per student and enough bottled water for 1 cup per student; 1 six-sided die and 1 index card per student

The ABC News program *20/20* set up a blind taste test in which people were asked to rate four different brands of bottled water and New York City tap water without knowing which they were drinking. Can you guess the result? Tap water came out the clear winner in terms of taste.

This Activity will give you and your classmates a chance to discover whether or not you can taste the difference between bottled water and tap water.

1. Before class begins, your teacher will prepare numbered stations with cups of water. You will be given an index card with a station number on it.
2. Go to the corresponding station. Pick up three cups (labeled A, B, and C) and take them back to your seat.
3. Your task is to determine which one of the three cups contains the bottled water. Drink all the water in Cup A first, then the water in Cup B, and finally the water in Cup C. Write down the letter of the cup that you think held the bottled water. Do not discuss your results with any of your classmates yet!
4. While you taste, your teacher will make a chart on the board like this one:

Station number	Bottled water cup?	Truth

5. When you are told to do so, go to the board and record your station number and the letter of the cup you identified as containing bottled water.
6. Your teacher will now reveal the truth about the cups of drinking water. How many students in the class identified the bottled water correctly? What percent of the class is this?
7. Let's assume that no one in your class can distinguish tap water from bottled water. In that case, students would just be guessing which cup of water tastes different. If so, what's the probability that an individual student would guess correctly?
8. How many correct identifications would you need to see to be convinced that the students in your class aren't just guessing? With your classmates, design and carry out a simulation to answer this question. What do you conclude about your class's ability to distinguish tap water from bottled water?

When Mr. Bullard's class did the preceding Activity, 13 out of 21 students made correct identifications. If we assume that the students in his class can't tell tap water from bottled water, then each one is basically guessing, with a 1/3 chance of being correct. So we'd expect about one-third of his 21 students, that is, about 7 students, to guess correctly. How likely is it that 13 or more of his 21 students would guess correctly? To answer this question without a simulation, we need a different kind of probability model from the ones we saw in Chapter 5.

Section 6.1 introduces the concept of a *random variable*, a numerical outcome of some chance process (like the 13 students who guessed correctly in Mr. Bullard's class). Each random variable has a *probability distribution* that gives us information about the likelihood that a specific event happens (like 13 or more correct guesses out of 21) and about what's expected to happen if the chance behavior is repeated many times. Section 6.2 examines the effect of transforming and combining random variables on the shape, center, and spread of their probability distributions. In Section 6.3, we'll look at two random variables with probability distributions that are used enough to have their own names—*binomial* and *geometric*.

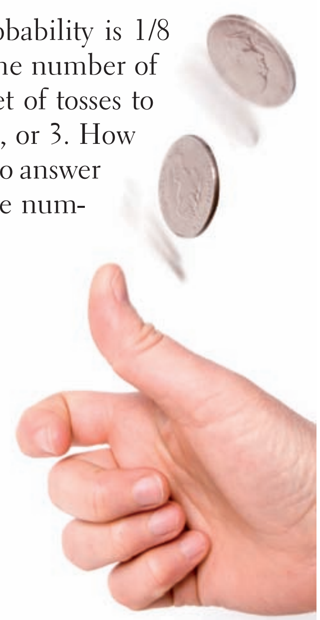
6.1 Discrete and Continuous Random Variables

WHAT YOU WILL LEARN **By the end of the section, you should be able to:**

- Compute probabilities using the probability distribution of a discrete random variable.
- Calculate and interpret the mean (expected value) of a discrete random variable.
- Calculate and interpret the standard deviation of a discrete random variable.
- Compute probabilities using the probability distribution of certain continuous random variables.

A probability model describes the possible outcomes of a chance process and the likelihood that those outcomes will occur. For example, suppose we toss a fair coin 3 times. The sample space for this chance process is

HHH HHT HTH THH HTT THT TTH TTT



Because there are 8 equally likely outcomes, the probability is 1/8 for each possible outcome. Define the variable X = the number of heads obtained. The value of X will vary from one set of tosses to another but will always be one of the numbers 0, 1, 2, or 3. How likely is X to take each of those values? It will be easier to answer this question if we group the possible outcomes by the number of heads obtained:

$X = 0$: TTT
$X = 1$: HTT THT TTH
$X = 2$: HHT HTH THH
$X = 3$: HHH

We can summarize the **probability distribution** of X as follows:

Value:	0	1	2	3
Probability:	1/8	3/8	3/8	1/8

Figure 6.1 on next page shows the probability distribution of X in graphical form. Notice the symmetric shape.

We can use the probability distribution to answer questions about the variable X.

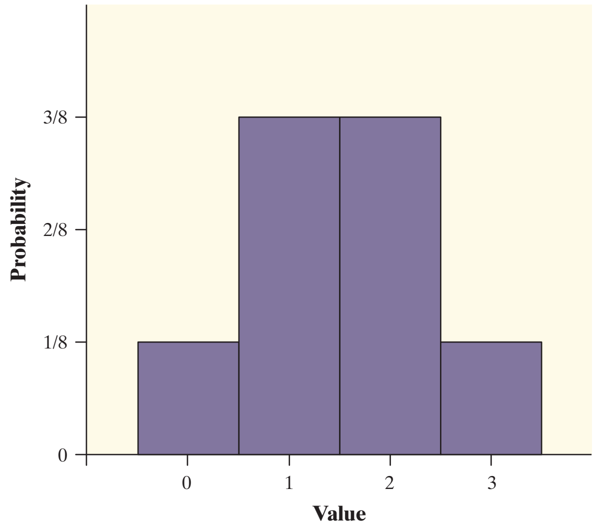


FIGURE 6.1 Histogram of the probability distribution for $X =$ number of heads in three tosses of a fair coin.

What's the probability that we get at least one head in three tosses of the coin? In symbols, we want to find $P(X \geq 1)$. We could add probabilities to get the answer:

$$P(X \geq 1) = P(X = 1) + P(X = 2) + P(X = 3)$$
$$= 1/8 + 3/8 + 3/8 = 7/8$$

Or we could use the complement rule from Chapter 5:

$$P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0)$$
$$= 1 - 1/8 = 7/8$$

A numerical variable that describes the outcomes of a chance process (like X in the coin-tossing scenario) is called a **random variable.** The probability model for a random variable is its probability distribution.

DEFINITION: Random variable and probability distribution

A **random variable** takes numerical values that describe the outcomes of some chance process. The **probability distribution** of a random variable gives its possible values and their probabilities.

There are two main types of random variables, corresponding to two types of probability distributions: *discrete* and *continuous*.

Discrete Random Variables

We have learned several rules of probability but only one way of assigning probabilities to events: assign a probability to every individual outcome, then add these probabilities to find the probability of any event. This idea works well if we can find a way to list all possible outcomes. We will call random variables having probability assigned in this way **discrete random variables.**[3] The probability distribution for a discrete random variable must have outcome probabilities that are between 0 and 1 and that add up to 1.

DISCRETE RANDOM VARIABLES AND THEIR PROBABILITY DISTRIBUTIONS

A **discrete random variable** X takes a fixed set of possible values with gaps between. The probability distribution of a discrete random variable X lists the values x_i and their probabilities p_i:

Value:	x_1	x_2	x_3	...
Probability:	p_1	p_2	p_3	...

The probabilities p_i must satisfy two requirements:

1. Every probability p_i is a number between 0 and 1.
2. The sum of the probabilities is 1: $p_1 + p_2 + p_3 + \cdots = 1$.

To find the probability of any event, add the probabilities p_i of the particular values x_i that make up the event.

The variable X in the coin-tossing example is a discrete random variable. We can list the possible values of X as 0, 1, 2, 3. Note that there are gaps between these values on a number line. The corresponding probabilities are all between 0 and 1, and their sum is 1/8 + 3/8 + 3/8 + 1/8 = 1.

Here's an example of a discrete random variable that involves something a bit more serious than tossing coins.

EXAMPLE Apgar Scores: Babies' Health at Birth

Discrete random variables

In 1952, Dr. Virginia Apgar suggested five criteria for measuring a baby's health at birth: skin color, heart rate, muscle tone, breathing, and response when stimulated. She developed a 0-1-2 scale to rate a newborn on each of the five criteria. A baby's Apgar score is the sum of the ratings on each of the five scales, which gives a whole-number value from 0 to 10. Apgar scores are still used today to evaluate the health of newborns.

What Apgar scores are typical? To find out, researchers recorded the Apgar scores of over 2 million newborn babies in a single year.[4] Imagine selecting one of these newborns at random. (That's our chance process.) Define the random variable X = Apgar score of a randomly selected baby one minute after birth. The table below gives the probability distribution for X.

Value:	0	1	2	3	4	5	6	7	8	9	10
Probability:	0.001	0.006	0.007	0.008	0.012	0.020	0.038	0.099	0.319	0.437	0.053

PROBLEM:

(a) Show that the probability distribution for X is legitimate.

(b) Make a histogram of the probability distribution. Describe what you see.

(c) Doctors decided that Apgar scores of 7 or higher indicate a healthy baby. What's the probability that a randomly selected baby is healthy?

SOLUTION:

(a) The probabilities are all between 0 and 1, and they add up to 1. So this is a legitimate probability distribution.

(b) Figure 6.2 shows a histogram of the probability distribution of X. *Shape:* The graph is skewed to the left and single-peaked. A randomly selected newborn will most likely have an Apgar score on the high end of the scale, which means that the baby was fairly healthy at birth. *Center:* From the probability distribution, we see that the median is 8. *Spread:* Apgar scores vary from 0 to 10. But most newborns receive scores between 4 and 10.

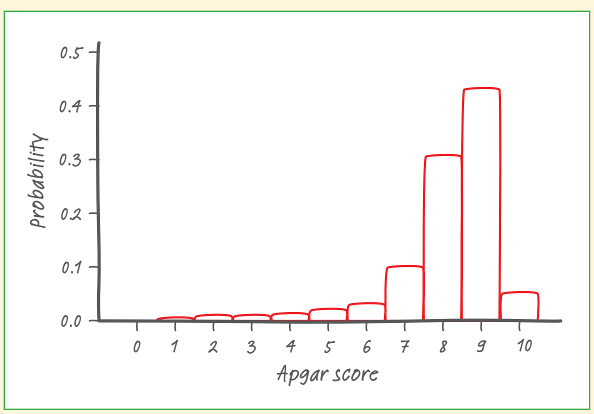


FIGURE 6.2 Histogram showing the probability distribution of the random variable X = Apgar score of a randomly selected newborn one minute after birth.

(c) The probability of choosing a healthy baby is $P(X \geq 7)$. We can calculate this probability as follows:

$$P(X \geq 7) = P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10)$$
$$= 0.099 + 0.319 + 0.437 + 0.053 = 0.908$$

That is, we'd have about a 91% chance of randomly choosing a healthy baby.

For Practice *Try Exercise* **5**

Although this procedure was later named for Dr. Apgar, the acronym APGAR also represents the five scales: Appearance, Pulse, Grimace, Activity, and Respiration.

Note that the probability of randomly selecting a newborn whose Apgar score is greater than or equal to 7 is not the same as the probability that the baby's Apgar score is strictly greater than 7. The latter probability is

$$P(X > 7) = P(X = 8) + P(X = 9) + P(X = 10)$$
$$= 0.319 + 0.437 + 0.053 = 0.809$$

The outcome $X = 7$ is included in "greater than or equal to" and is not included in "greater than." Be sure to confirm the values of interest when dealing with discrete random variables.

CHECK YOUR UNDERSTANDING

North Carolina State University posts the grade distributions for its courses online.[5] Students in Statistics 101 in a recent semester received 26% A's, 42% B's, 20% C's, 10% D's, and 2% F's. Choose a Statistics 101 student at random. The student's grade on a four-point scale (with A = 4) is a discrete random variable X with this probability distribution:

Value of X:	0	1	2	3	4
Probability:	0.02	0.10	0.20	0.42	0.26

1. Say in words what the meaning of $P(X \geq 3)$ is. What is this probability?
2. Write the event "the student got a grade worse than C" in terms of values of the random variable X. What is the probability of this event?
3. Sketch a graph of the probability distribution. Describe what you see.

Mean (Expected Value) of a Discrete Random Variable

When we analyzed distributions of quantitative data in Chapter 1, we made it a point to discuss their shape, center, and spread. We'll follow the same strategy with probability distributions of random variables. You can use what you learned earlier to describe the shape of a probability distribution histogram. We've already seen examples of symmetric (number of heads in three coin tosses) and left-skewed (Apgar score of a randomly chosen baby) probability distributions. What about center and spread?

The mean $\bar{x}$ of a set of observations is their average. The **mean μ_X of a discrete random variable** X is also an average of the possible values of X but with an important change to take into account the fact that not all outcomes may be equally likely. A simple example will show what we need to do.

EXAMPLE Winning (and Losing) at Roulette

Finding the mean of a discrete random variable

On an American roulette wheel, there are 38 slots numbered 1 through 36, plus 0 and 00. Half of the slots from 1 to 36 are red; the other half are black. Both the 0 and 00 slots are green. Suppose that a player places a simple \$1 bet on red. If the ball lands in a red slot, the player gets the original dollar back, plus an additional

dollar for winning the bet. If the ball lands in a different-colored slot, the player loses the dollar bet to the casino.

Let's define the random variable X = net gain from a single \$1 bet on red. The possible values of X are −\$1 and \$1. (The player either gains a dollar or loses a dollar.) What are the corresponding probabilities? The chance that the ball lands in a red slot is 18/38. The chance that the ball lands in a different-colored slot is 20/38. Here is the probability distribution of X:

Value:	−\$1	\$1
Probability:	20/38	18/38

What is the player's average gain? The ordinary average of the two possible outcomes −\$1 and \$1 is \$0. But \$0 isn't the average winnings because the player is less likely to win \$1 than to lose \$1. In the long run, the player gains a dollar 18 times in every 38 games played and loses a dollar on the remaining 20 of 38 bets. The player's long-run average gain for this simple bet is

$$\mu_X = (-\$1)\left(\frac{20}{38}\right) + (\$1)\left(\frac{18}{38}\right) = -\$0.05$$

You see that the player loses (and the casino gains) an average of five cents per \$1 bet in many, many plays of the game.

If someone played several games of roulette, we would call the mean amount the person gained $\bar{x}$. The mean in the previous example is a different quantity—it is the long-run average gain we'd expect if someone played roulette a very large number of times. For this reason, the mean of a random variable is often referred to as its **expected value.** Just as probabilities describe the proportion of times that an outcome occurs in many repetitions of a chance process, the mean of a discrete random variable describes the long-run average outcome.

There are two ways of denoting the mean of a random variable X. We can use the notation μ_X, or we can write $E(X)$, as in the "expected value of X." In the roulette example, $\mu_X = E(X) = -\$0.05$.

The mean of any discrete random variable is found just as in the roulette example. It is an average of the possible outcomes, but a weighted average in which each outcome is weighted by its probability. Here (finally!) is the definition.

DEFINITION: Mean (expected value) of a discrete random variable

Suppose that X is a discrete random variable with probability distribution

Value:	x_1	x_2	x_3	...
Probability:	p_1	p_2	p_3	...

To find the **mean (expected value)** of X, multiply each possible value by its probability, then add all the products:

$$\mu_X = E(X) = x_1p_1 + x_2p_2 + x_3p_3 + \ldots$$
$$= \sum x_i p_i$$

Let's put the definition to use in calculating the mean of a familiar random variable.

EXAMPLE Apgar Scores: What's Typical?

Mean and expected value as an average

In our earlier example, we defined the random variable X to be the Apgar score of a randomly selected baby. The table below gives the probability distribution for X once again.

Value x_i:	0	1	2	3	4	5	6	7	8	9	10
Probability p_i:	0.001	0.006	0.007	0.008	0.012	0.020	0.038	0.099	0.319	0.437	0.053

PROBLEM: Compute the mean of the random variable X. Interpret this value in context.

SOLUTION: From the probability distribution for X, we see that 1 in every 1000 babies would have an Apgar score of 0; 6 in every 1000 babies would have an Apgar score of 1; and so on. So the mean (expected value) of X is

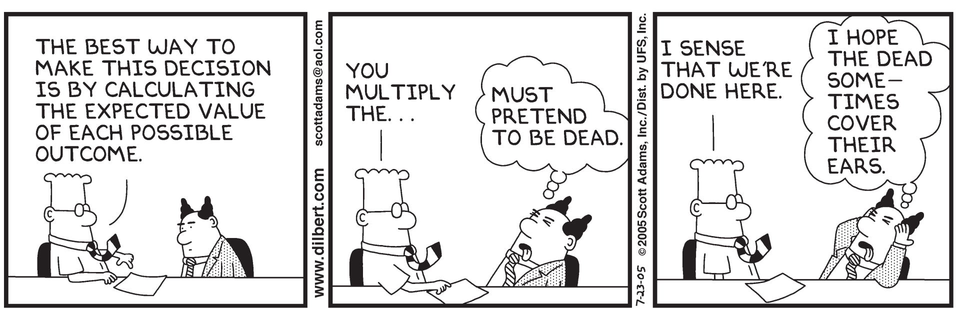
$$\mu_X = E(X) = \sum x_i p_i$$
$$= (0)(0.001) + (1)(0.006) + (2)(0.007) + \cdots + (10)(0.053) = 8.128$$

The mean Apgar score of a randomly selected newborn is 8.128. This is the average Apgar score of many, many randomly chosen babies.

For Practice *Try Exercise* 9

AP® EXAM TIP If the mean of a random variable has a non-integer value, but you report it as an integer, your answer will not get full credit.

Notice that the mean Apgar score, 8.128, is not a possible value of the random variable X. It's also not an integer. If you think of the mean as a long-run average over many repetitions, these facts shouldn't bother you.



Standard Deviation (and Variance) of a Discrete Random Variable

With the mean as our measure of center for a discrete random variable, it shouldn't surprise you that we'll use the standard deviation as our measure of spread. In Chapter 1, we first defined the sample variance s_x^2 as the "typical" squared deviation from the mean and then took the square root of the variance to get the sample standard deviation s_x. The definition of the **variance of a random variable** σ_X^2 is similar to the definition of the variance for a set of quantitative data. That is,

Recall that the formula for the sample variance is

$$s_x^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

the variance is an "average" of the squared deviation $(x_i - \mu_X)^2$ of the values of the variable X from its mean μ_X. As with the mean, the average we use is a weighted average. Each outcome is weighted by its probability to take account of outcomes that are not equally likely. To get the **standard deviation of a random variable,** we take the square root of the variance. Here are the details.

DEFINITION: Variance and standard deviation of a discrete random variable

Suppose that X is a discrete random variable with probability distribution

Value:	x_1	x_2	x_3	...
Probability:	p_1	p_2	p_3	...

and that μ_X is the mean of X. The **variance** of X is

$$\begin{aligned}\text{Var}(X) = \sigma_X^2 &= (x_1 - \mu_X)^2 p_1 + (x_2 - \mu_X)^2 p_2 + (x_3 - \mu_X)^2 p_3 + \ldots \\ &= \sum(x_i - \mu_X)^2 p_i\end{aligned}$$

The **standard deviation** of X, σ_X, is the square root of the variance.

$$\sigma_X = \sqrt{\sum(x_i - \mu_X)^2 p_i}$$

The standard deviation of a random variable X is a measure of how much the values of the variable typically vary from the mean μ_X. Let's compute the variance and standard deviation of a familiar discrete random variable.

EXAMPLE Apgar Scores: How Variable Are They?

Calculating measures of spread

In the last example, we calculated the mean Apgar score of a randomly chosen newborn to be $\mu_X = 8.128$. The table below gives the probability distribution for X one more time.

Value x_i:	0	1	2	3	4	5	6	7	8	9	10
Probability p_i:	0.001	0.006	0.007	0.008	0.012	0.020	0.038	0.099	0.319	0.437	0.053

PROBLEM: *Compute and interpret the standard deviation of the random variable X.*

SOLUTION: The formula for the variance of X is $\sigma_X^2 = \sum(x_i - \mu_X)^2 p_i$. Plugging in values gives

$$\begin{aligned}\sigma_X^2 &= (0 - 8.128)^2(0.001) + (1 - 8.128)^2(0.006) \\ &\quad + (2 - 8.128)^2(0.007) + \cdots + (10 - 8.128)^2(0.053) \\ \sigma_X^2 &= 2.066\end{aligned}$$

The standard deviation of X is $\sigma_X = \sqrt{2.066} = 1.437$. A randomly selected baby's Apgar score will typically differ from the mean (8.128) by about 1.4 units.

For Practice *Try Exercise* **15**

You can use your calculator to graph the probability distribution of a discrete random variable and to calculate measures of center and spread, as the following Technology Corner illustrates.

11. TECHNOLOGY CORNER

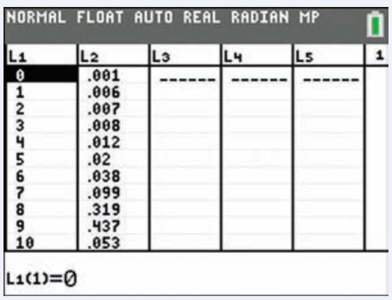
ANALYZING RANDOM VARIABLES ON THE CALCULATOR

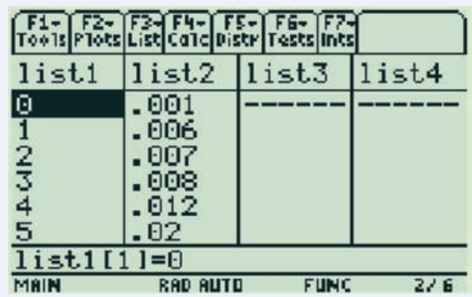
TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

Let's explore what the calculator can do using the random variable X = Apgar score of a randomly selected newborn.

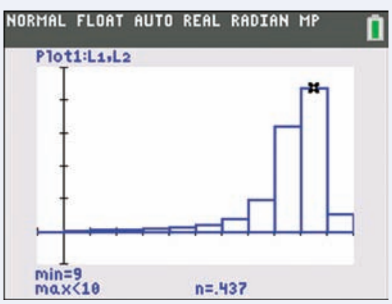
TI-83/84 **TI-89**

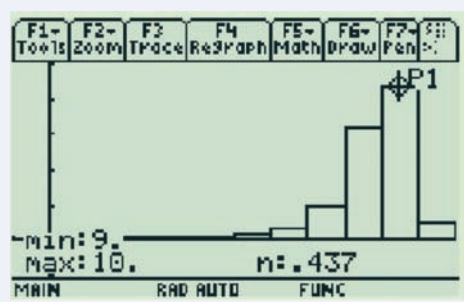
1. Start by entering the values of the random variable in L1/list1 and the corresponding probabilities in L2/list2.





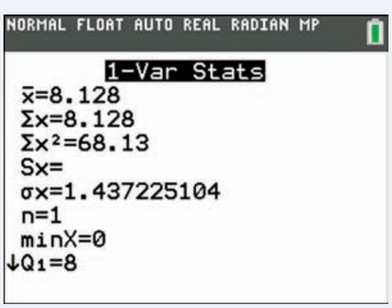
2. To graph a histogram of the probability distribution:
   - Set up a statistics plot with Xlist: L1/list1 and Freq: L2/list2.
   - Adjust your window settings as follows: Xmin = −1, Xmax = 11, Xscl = 1,Ymin = −0.1, Ymax = 0.5, Yscl = 0.1.
   - Press GRAPH (◆ F3 on the TI-89).





3. To calculate the mean and standard deviation of the random variable, use one-variable statistics with the values in L1/list1 and the probabilities (relative frequencies) in L2/list2.

   - **OS 2.55 or later:** In the dialog box, specify List: L1 and FreqList: L2. Then choose Calculate. **Older OS:** Execute the command `1-Var Stats L1,L2`.
   - In the Statistics/List Editor, press F4 (Calc) and choose `1-Var Stats...` Use the inputs List: list1 and Freq: list2.



CHECK YOUR UNDERSTANDING

A large auto dealership keeps track of sales made during each hour of the day. Let X = the number of cars sold during the first hour of business on a randomly selected Friday. Based on previous records, the probability distribution of X is as follows:

Cars sold:	0	1	2	3
Probability:	0.3	0.4	0.2	0.1

1. Compute and interpret the mean of X.
2. Compute and interpret the standard deviation of X.

Continuous Random Variables

When we use the table of random digits to select a digit between 0 and 9, the result is a discrete random variable (call it X). The probability model assigns probability 1/10 to each of the 10 possible values of X.

The calculator command `rand` will generate a random number from 0 to 1. Can you figure out how to modify the command to find a random number between, say, 1 and 3?

Suppose we want to choose a number at random between 0 and 1, allowing *any* number between 0 and 1 as the outcome (like 0.84522 or 0.1111119). Calculator and computer random number generators will do this. The sample space of this chance process is an entire interval of numbers:

$$S = \text{all numbers between 0 and 1}$$

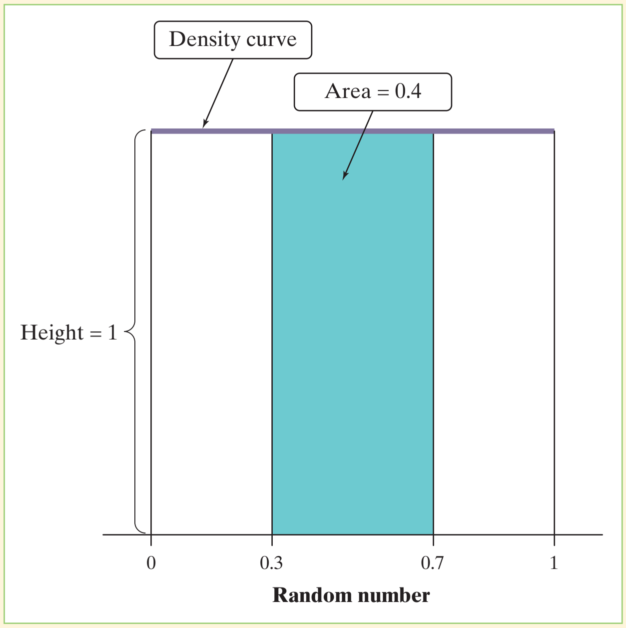
Call the outcome of the random number generator Y for short. How can we find probabilities of events like $P(0.3 \leq Y \leq 0.7)$? As in the case of selecting a random digit, we would like all possible outcomes to be equally likely. But we cannot assign probabilities to each individual value of Y and then add them, because there are infinitely many possible values.

In situations like this, we use a different way of assigning probabilities directly to events—as *areas under a density curve*. Recall from Chapter 2 that any density curve has area exactly 1 underneath it, corresponding to total probability 1.

EXAMPLE Random Numbers

Density curves and probability distributions

The random number generator will spread its output uniformly across the entire interval from 0 to 1 as we allow it to generate a long sequence of random numbers. The results of many trials are represented by the density curve of a *uniform distribution*. This density curve appears in purple in Figure 6.3 on the next page. It has height 1 over the interval from 0 to 1. The area under the density curve is 1, and the probability of any event is the area under the density curve and above the event in question.



As Figure 6.3 shows, the probability that the random number generator produces a number Y between 0.3 and 0.7 is

$$P(0.3 \leq Y \leq 0.7) = 0.4$$

That's because the area of the shaded rectangle is

$$\text{length} \times \text{width} = 0.4 \times 1 = 0.4$$

FIGURE 6.3 Assigning probabilities for generating a random number between 0 and 1. The probability of any interval of numbers is the area above the interval and under the density curve. The shaded area represents $P(0.3 \leq Y \leq 0.7)$

In many cases, discrete random variables arise from counting something—for instance, the number of siblings that a randomly selected student has. Continuous random variables often arise from measuring something—for instance, the height or time to run a mile for a randomly selected student.

Figure 6.3 shows the probability distribution of the random variable Y = random number between 0 and 1. We call Y a **continuous random variable** because its values are not isolated numbers but rather an entire interval of numbers.

DEFINITION: Continuous random variable

A **continuous random variable** X takes all values in an interval of numbers. The probability distribution of X is described by a density curve. The probability of any event is the area under the density curve and above the values of X that make up the event.

The probability distribution for a continuous random variable assigns probabilities to intervals of outcomes rather than to individual outcomes. In fact, *all continuous probability models assign probability 0 to every individual outcome.* Only intervals of values have positive probability. To see that this is true, consider a specific outcome from the random number generator of the previous example, such as $P(Y = 0.7)$. The probability of this event is the area under the density curve that's above the point 0.70000 … on the horizontal axis. But this vertical line segment has no width, so the area is 0. For that reason,

$$P(0.3 \leq Y \leq 0.7) = P(0.3 \leq Y < 0.7) = P(0.3 < Y < 0.7) = 0.4$$

We can use any density curve to assign probabilities. The density curves that are most familiar to us are the Normal curves of Chapter 2. We learned how to find areas in any Normal distribution on page 118. Normal distributions can be probability distributions as well as models for data. The following example shows the connection between the two.

EXAMPLE Young Women's Heights

Normal probability distributions

The heights of young women closely follow the Normal distribution with mean $\mu = 64$ inches and standard deviation $\sigma = 2.7$ inches. This is a distribution for a large set of data. Now choose one young woman at random. Call her height Y. If we repeat the random choice very many times, the distribution of values of Y is the same Normal distribution that describes the heights of all young women. Find the probability that the chosen woman is between 68 and 70 inches tall.

PROBLEM: What's the probability that a randomly chosen young woman has height between 68 and 70 inches?

SOLUTION:

Step 1: State the distribution and the values of interest. The height Y of a randomly chosen young woman has the $N(64, 2.7)$ distribution. We want to find $P(68 \le Y \le 70)$. Figure 6.4 shows the distribution with the area of interest shaded and the mean, standard deviation, and boundary values labeled.

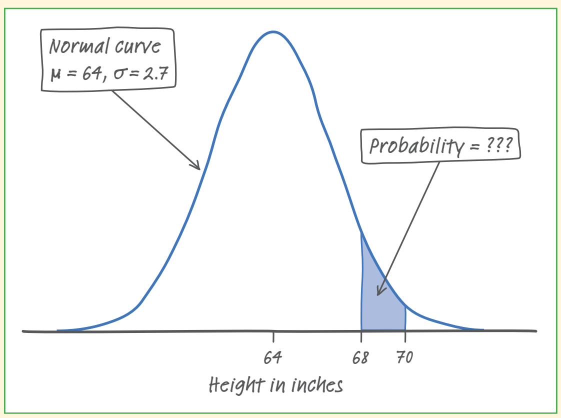


FIGURE 6.4 The probability that a randomly chosen young woman has height between 68 and 70 inches as an area under a Normal curve.

AP® EXAM TIP When showing your work on a free response question, you must include more than a calculator command. Writing `normalcdf(68,70,64,2.7)` will *not* earn you full credit for a Normal calculation. At a minimum, you must indicate what each of those calculator inputs represents. Better yet, sketch and label a Normal curve to show what you're finding.

Step 2: Perform calculations—show your work! The standardized scores for the two boundary values are

$$z = \frac{68 - 64}{2.7} = 1.48 \text{ and } z = \frac{70 - 64}{2.7} = 2.22$$

The random variable Z follows a standard Normal distribution, and the desired probability is $P(1.48 \le Z \le 2.22)$. From Table A, we find that $P(Z \le 2.22) = 0.9868$ and $P(Z \le 1.48) = 0.9306$. So we have

$$P(1.48 \le Z \le 2.22) = P(Z \le 2.22) - P(Z \le 1.48)$$
$$= 0.9868 - 0.9306 = 0.0562$$

Using technology: The command `normalcdf(lower:68, upper:70, μ:64, σ:2.7)` gives an area of 0.0561.

Step 3: Answer the question. There's about a 5.6% chance that a randomly chosen young woman has a height between 68 and 70 inches.

For Practice *Try Exercise* **23**

The calculation in the preceding example is the same as those we did in Chapter 2. Only the language of probability is new.

What about the mean and standard deviation for continuous random variables? The probability distribution of a continuous random variable X is described by a density curve. Chapter 2 showed how to find the mean of the distribution: it is the point at which the area under the density curve would balance if it were made out of solid material. The mean lies at the center of symmetric density curves such as the Normal curves. We can locate the standard deviation of a Normal distribution from its inflection points. Exact calculation of the mean and standard deviation for most continuous random variables requires advanced mathematics.[6]

Section 6.1 Summary

- A **random variable** takes numerical values determined by the outcome of a chance process. The **probability distribution** of a random variable X tells us what the possible values of X are and how probabilities are assigned to those values. There are two types of random variables: *discrete* and *continuous*.
- A **discrete random variable** has a fixed set of possible values with gaps between them. The probability distribution assigns each of these values a probability between 0 and 1 such that the sum of all the probabilities is exactly 1. The probability of any event is the sum of the probabilities of all the values that make up the event.
- A **continuous random variable** takes all values in some interval of numbers. A density curve describes the probability distribution of a continuous random variable. The probability of any event is the area under the curve above the values that make up the event.
- The **mean of a random variable** μ_X is the balance point of the probability distribution histogram or density curve. Because the mean is the long-run average value of the variable after many repetitions of the chance process, it is also known as the **expected value** of the random variable, $E(X)$.
- If X is a discrete random variable, the mean is the average of the values of X, each weighted by its probability:

$$\mu_X = E(X) = \sum x_i p_i = x_1 p_1 + x_2 p_2 + x_3 p_3 + \ldots$$

- The **variance of a random variable** σ_X^2 is the "average" squared deviation of the values of the variable from their mean. The **standard deviation** σ_X is the square root of the variance. The standard deviation measures the typical distance of the values in the distribution from the mean.
- For a discrete random variable X, the variance is

$$\sigma_X^2 = \sum (x_i - \mu_X)^2 p_i = (x_1 - \mu_X)^2 p_1 + (x_2 - \mu_X)^2 p_2 + (x_3 - \mu_X)^2 p_3 + \cdots$$

and the standard deviation is

$$\sigma_X = \sqrt{\sum (x_i - \mu_X)^2 p_i}$$

6.1 TECHNOLOGY CORNER

TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

Section 6.1 Exercises

1. **Toss 4 times** Suppose you toss a fair coin 4 times. Let X = the number of heads you get.

(a) Find the probability distribution of X.

(b) Make a histogram of the probability distribution. Describe what you see.

(c) Find $P(X \leq 3)$ and interpret the result.

2. **Pair-a-dice** Suppose you roll a pair of fair, six-sided dice. Let T = the sum of the spots showing on the up-faces.

(a) Find the probability distribution of T.

(b) Make a histogram of the probability distribution. Describe what you see.

(c) Find $P(T \geq 5)$ and interpret the result.

3. **Spell-checking** Spell-checking software catches "nonword errors," which result in a string of letters that is not a word, as when "the" is typed as "teh." When undergraduates are asked to write a 250-word essay (without spell-checking), the number X of nonword errors has the following distribution:

Value:	0	1	2	3	4
Probability:	0.1	0.2	0.3	0.3	0.1

(a) Write the event "at least one nonword error" in terms of X. What is the probability of this event?

(b) Describe the event $X \leq 2$ in words. What is its probability? What is the probability that $X < 2$?

4. **Kids and toys** In an experiment on the behavior of young children, each subject is placed in an area with five toys. Past experiments have shown that the probability distribution of the number X of toys played with by a randomly selected subject is as follows:

Number of toys x_i:	0	1	2	3	4	5
Probability p_i:	0.03	0.16	0.30	0.23	0.17	0.11

(a) Write the event "plays with at most two toys" in terms of X. What is the probability of this event?

(b) Describe the event $X > 3$ in words. What is its probability? What is the probability that $X \geq 3$?

5. pg 349 **Benford's law** Faked numbers in tax returns, invoices, or expense account claims often display patterns that aren't present in legitimate records. Some patterns, like too many round numbers, are obvious and easily avoided by a clever crook. Others are more subtle. It is a striking fact that the first digits of numbers in legitimate records often follow a model known as Benford's law.[7] Call the first digit of a randomly chosen record X for short. Benford's law gives this probability model for X (note that a first digit can't be 0):

First digit:	1	2	3	4	5	6	7	8	9
Probability:	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

(a) Show that this is a legitimate probability distribution.

(b) Make a histogram of the probability distribution. Describe what you see.

(c) Describe the event $X \geq 6$ in words. What is $P(X \geq 6)$?

(d) Express the event "first digit is at most 5" in terms of X. What is the probability of this event?

6. **Working out** Choose a person aged 19 to 25 years at random and ask, "In the past seven days, how many times did you go to an exercise or fitness center or work out?" Call the response Y for short. Based on a large sample survey, here is a probability model for the answer you will get:[8]

Days:	0	1	2	3	4	5	6	7
Probability:	0.68	0.05	0.07	0.08	0.05	0.04	0.01	0.02

(a) Show that this is a legitimate probability distribution.

(b) Make a histogram of the probability distribution. Describe what you see.

(c) Describe the event $Y < 7$ in words. What is $P(Y < 7)$?

(d) Express the event "worked out at least once" in terms of Y. What is the probability of this event?

7. **Benford's law** Refer to Exercise 5. The first digit of a randomly chosen expense account claim follows Benford's law. Consider the events A = first digit is 7 or greater and B = first digit is odd.

(a) What outcomes make up the event A? What is $P(A)$?

(b) What outcomes make up the event B? What is $P(B)$?

(c) What outcomes make up the event "A or B"? What is $P(A \text{ or } B)$? Why is this probability not equal to $P(A) + P(B)$?

8. **Working out** Refer to Exercise 6. Consider the events A = works out at least once and B = works out less than 5 times per week.

(a) What outcomes make up the event A? What is $P(A)$?

(b) What outcomes make up the event B? What is $P(B)$?

(c) What outcomes make up the event "A and B"? What is $P(A \text{ and } B)$? Why is this probability not equal to $P(A) \cdot P(B)$?

pg 352

9. **Keno** Keno is a favorite game in casinos, and similar games are popular with the states that operate lotteries. Balls numbered 1 to 80 are tumbled in a machine as the bets are placed, then 20 of the balls are chosen at random. Players select numbers by marking a card. The simplest of the many wagers available is "Mark 1 Number." Your payoff is \$3 on a \$1 bet if the number you select is one of those chosen. Because 20 of 80 numbers are chosen, your probability of winning is 20/80, or 0.25. Let X = the net amount you gain on a single play of the game.

(a) Make a table that shows the probability distribution of X.

(b) Compute the expected value of X. Explain what this result means for the player.

10. **Fire insurance** Suppose a homeowner spends \$300 for a home insurance policy that will pay out \$200,000 if the home is destroyed by fire. Let Y = the profit made by the company on a single policy. From previous data, the probability that a home in this area will be destroyed by fire is 0.0002.

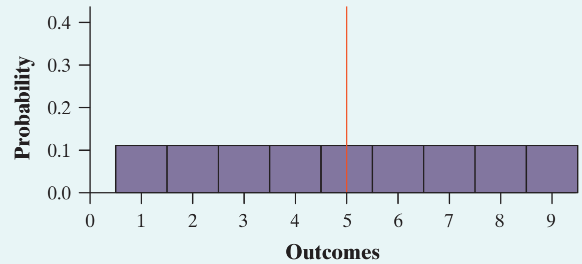
(a) Make a table that shows the probability distribution of Y.

(b) Compute the expected value of Y. Explain what this result means for the insurance company.

11. **Spell-checking** Refer to Exercise 3. Calculate the mean of the random variable X and interpret this result in context.

12. **Kids and toys** Refer to Exercise 4. Calculate the mean of the random variable X and interpret this result in context.

13. **Benford's law and fraud** A not-so-clever employee decided to fake his monthly expense report. He believed that the first digits of his expense amounts should be equally likely to be any of the numbers from 1 to 9. In that case, the first digit Y of a randomly selected expense amount would have the probability distribution shown in the histogram.



(a) Explain why the mean of the random variable Y is located at the solid red line in the figure.

(b) The first digits of randomly selected expense amounts actually follow Benford's law (Exercise 5). According to Benford's law, what's the expected value of the first digit? Explain how this information could be used to detect a fake expense report.

(c) What's $P(Y > 6)$ in the above distribution? According to Benford's law, what proportion of first digits in the employee's expense amounts should be greater than 6? How could this information be used to detect a fake expense report?

14. **Life insurance** A life insurance company sells a term insurance policy to a 21-year-old male that pays \$100,000 if the insured dies within the next 5 years. The probability that a randomly chosen male will die each year can be found in mortality tables. The company collects a premium of \$250 each year as payment for the insurance. The amount Y that the company earns on this policy is \$250 per year, less the \$100,000 that it must pay if the insured dies. Here is a partially completed table that shows information about risk of mortality and the values of Y = profit earned by the company:

Age at death:	21	22	23	24	25	26 or more
Profit:	−\$99,750	−\$99,500	−\$99,250	−\$99,000	−\$98,750	\$1250
Probability:	0.00183	0.00186	0.00189	0.00191	0.00193	

(a) Explain why the company suffers a loss of \$98,750 on such a policy if a client dies at age 25.

(b) Find the missing probability. Show your work.

(c) Calculate the mean μ_Y. Interpret this value in context.

15. **Spell-checking** Refer to Exercise 3. Calculate and interpret the standard deviation of the random variable X. Show your work.
pg 353

16. **Kids and toys** Refer to Exercise 4. Calculate and interpret the standard deviation of the random variable X. Show your work.

17. **Benford's law and fraud** Refer to Exercise 13. It might also be possible to detect an employee's fake expense records by looking at the variability in the first digits of those expense amounts.

(a) Calculate the standard deviation σ_Y. This gives us an idea of how much variation we'd expect in the employee's expense records if he assumed that first digits from 1 to 9 were equally likely.

(b) Now calculate the standard deviation of first digits that follow Benford's law (Exercise 5). Would using standard deviations be a good way to detect fraud? Explain.

18. **Life insurance**

(a) It would be quite risky for you to insure the life of a 21-year-old friend under the terms of Exercise 14. There is a high probability that your friend would live and you would gain \$1250 in premiums. But if he were to die, you would lose almost \$100,000. Explain carefully why selling insurance is not risky for an insurance company that insures many thousands of 21-year-old men.

(b) The risk of an investment is often measured by the standard deviation of the return on the investment. The more variable the return is, the riskier the investment. We can measure the great risk of insuring a single person's life in Exercise 14 by computing the standard deviation of the income Y that the insurer will receive. Find σ_Y using the distribution and mean found in Exercise 14.

19. **Housing in San Jose** How do rented housing units differ from units occupied by their owners? Here are the distributions of the number of rooms for owner-occupied units and renter-occupied units in San Jose, California:[9]

	Number of Rooms									
	1	2	3	4	5	6	7	8	9	10
Owned	0.003	0.002	0.023	0.104	0.210	0.224	0.197	0.149	0.053	0.035
Rented	0.008	0.027	0.287	0.363	0.164	0.093	0.039	0.013	0.003	0.003

Let X = the number of rooms in a randomly selected owner-occupied unit and Y = the number of rooms in a randomly chosen renter-occupied unit.

(a) Make histograms suitable for comparing the probability distributions of X and Y. Describe any differences that you observe.

(b) Find the mean number of rooms for both types of housing unit. Explain why this difference makes sense.

(c) Find and interpret the standard deviations of both X and Y.

20. **Size of American households** In government data, a household consists of all occupants of a dwelling unit, while a family consists of two or more persons who live together and are related by blood or marriage. So all families form households, but some households are not families. Here are the distributions of household size and family size in the United States:

	Number of Persons						
	1	2	3	4	5	6	7
Household probability	0.25	0.32	0.17	0.15	0.07	0.03	0.01
Family probability	0	0.42	0.23	0.21	0.09	0.03	0.02

Let X = the number of people in a randomly selected U.S. household and Y = the number of people in a randomly chosen U.S. family.

(a) Make histograms suitable for comparing the probability distributions of X and Y. Describe any differences that you observe.

(b) Find the mean for each random variable. Explain why this difference makes sense.

(c) Find and interpret the standard deviations of both X and Y.

21. **Random numbers** Let X be a number between 0 and 1 produced by a random number generator. Assuming that the random variable X has a uniform distribution, find the following probabilities:

(a) $P(X > 0.49)$

(b) $P(X \geq 0.49)$

(c) $P(0.19 \leq X < 0.37 \text{ or } 0.84 < X \leq 1.27)$

22. **Random numbers** Let Y be a number between 0 and 1 produced by a random number generator. Assuming that the random variable Y has a uniform distribution, find the following probabilities:

(a) $P(Y \leq 0.4)$

(b) $P(Y < 0.4)$

(c) $P(0.1 < Y \leq 0.15 \text{ or } 0.77 \leq Y < 0.88)$

23. **Running a mile** A study of 12,000 able-bodied male students at the University of Illinois found that their times for the mile run were approximately Normal with mean 7.11 minutes and standard deviation 0.74 minute.[10] Choose a student at random from this group and call his time for the mile Y. Find $P(Y < 6)$ and interpret the result.
pg 357

24. **ITBS scores** The Normal distribution with mean $\mu = 6.8$ and standard deviation $\sigma = 1.6$ is a good description of the Iowa Test of Basic Skills (ITBS) vocabulary scores of seventh-grade students in Gary, Indiana. Call the score of a randomly chosen student X for short. Find $P(X \geq 9)$ and interpret the result.

25. **Ace!** Professional tennis player Rafael Nadal hits the ball extremely hard. His first-serve speeds follow a Normal distribution with mean 115 miles per hour (mph) and standard deviation 6 mph. Choose one of Nadal's first serves at random. Let Y = its speed, measured in miles per hour.

(a) Find $P(Y > 120)$ and interpret the result.

(b) What is $P(Y \geq 120)$? Explain.

(c) Find the value of c such that $P(Y \leq c) = 0.15$. Show your work.

26. **Pregnancy length** The length of human pregnancies from conception to birth follows a Normal distribution with mean 266 days and standard deviation 16 days. Choose a pregnant woman at random. Let X = the length of her pregnancy.

(a) Find $P(X \geq 240)$ and interpret the result.

(b) What is $P(X > 240)$? Explain.

(c) Find the value of c such that $P(X \geq c) = 0.20$. Show your work.

Multiple choice: Select the best answer for Exercises 27 to 30.

Exercises 27 to 29 refer to the following setting. Choose an American household at random and let the random variable X be the number of cars (including SUVs and light trucks) they own. Here is the probability model if we ignore the few households that own more than 5 cars:

Number of cars X:	0	1	2	3	4	5
Probability:	0.09	0.36	0.35	0.13	0.05	0.02

27. What's the expected number of cars in a randomly selected American household?

(a) 1.00 (b) 1.75 (c) 1.84 (d) 2.00 (e) 2.50

28. The standard deviation of X is $\sigma_X = 1.08$. If many households were selected at random, which of the following would be the best interpretation of the value 1.08?

(a) The mean number of cars would be about 1.08.

(b) The number of cars would typically be about 1.08 from the mean.

(c) The number of cars would be at most 1.08 from the mean.

(d) The number of cars would be within 1.08 from the mean about 68% of the time.

(e) The mean number of cars would be about 1.08 from the expected value.

29. About what percentage of households have a number of cars within 2 standard deviations of the mean?

(a) 68% (b) 71% (c) 93% (d) 95% (e) 98%

30. A deck of cards contains 52 cards, of which 4 are aces. You are offered the following wager: Draw one card at random from the deck. You win \$10 if the card drawn is an ace. Otherwise, you lose \$1. If you make this wager very many times, what will be the mean amount you win?

(a) About −\$1, because you will lose most of the time.

(b) About \$9, because you win \$10 but lose only \$1.

(c) About −\$0.15; that is, on average you lose about 15 cents.

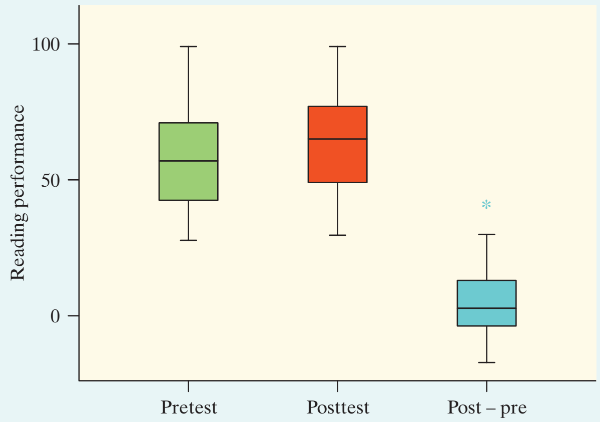
(d) About \$0.77; that is, on average you win about 77 cents.

(e) About \$0, because the random draw gives you a fair bet.

Exercises 31 to 34 refer to the following setting. Many chess masters and chess advocates believe that chess play develops general intelligence, analytical skill, and the ability to concentrate. According to such beliefs, improved reading skills should result from study to improve chess-playing skills. To investigate this belief, researchers conducted a study. All of the subjects in the study participated in a comprehensive chess program, and their reading performances were measured before and after the program. The graphs and numerical summaries below provide information on the subjects' pretest scores, posttest scores, and the difference (post – pre) between these two scores.

Descriptive Statistics: Pretest, Posttest, Post – pre

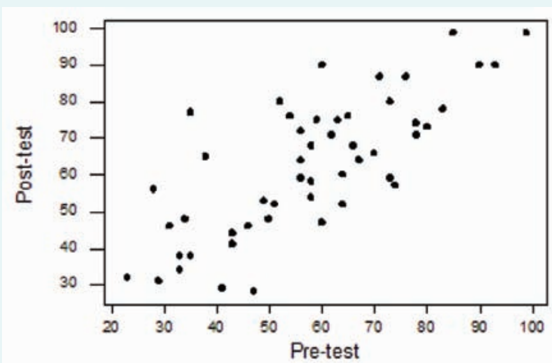
Variable	N	Mean	Median	StDev	Min	Max	Q_1	Q_3
Pretest	53	57.70	58.00	17.84	23.00	99.00	44.50	70.50
Posttest	53	63.08	64.00	18.70	28.00	99.00	48.00	76.00
Post-pre	53	5.38	3.00	13.02	-19.00	42.00	-3.50	14.00

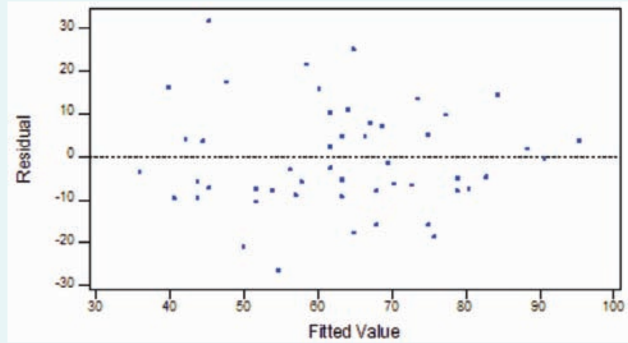


31. **Better readers?** (1.3) Did students have higher reading scores after participating in the chess program? Give appropriate statistical evidence to support your answer.

32. **Chess and reading** (4.3) If the study found a statistically significant improvement in reading scores, could you conclude that playing chess causes an increase in reading skills? Justify your answer.

Some graphical and numerical information about the relationship between pretest and posttest scores is provided below.





Regression Analysis: Posttest versus Pretest

Predictor	Coef	SE Coef	T	P
Constant	17.897	5.889	3.04	0.004
Pretest	0.78301	0.09758	8.02	0.000
S = 12.55	R-Sq = 55.8%		R-Sq(adj) = 54.9%	

33. **Predicting posttest scores** (3.2) What is the equation of the linear regression model relating posttest and pretest scores? Define any variables used.

34. **How well does it fit?** (3.2) Discuss what s, r^2, and the residual plot tell you about this linear regression model.

6.2 Transforming and Combining Random Variables

WHAT YOU WILL LEARN

By the end of the section, you should be able to:

- Describe the effects of transforming a random variable by adding or subtracting a constant and multiplying or dividing by a constant.
- Find the mean and standard deviation of the sum or difference of independent random variables.
- Find probabilities involving the sum or difference of independent Normal random variables.

In Section 6.1, we looked at several examples of random variables and their probability distributions. We also saw that the mean μ_X and standard deviation σ_X give us important information about a random variable. For instance, for X = the amount gained on a single \$1 bet on red in a game of roulette, we already showed that $\mu_X = -\$0.05$. You can verify that the standard deviation is $\sigma_X = \$1.00$. That is, a player can expect to lose an average of 5 cents per \$1 bet if he plays many games. But if he plays only a few games, his actual gain could be much better or worse than this expected value.

Would the player be better off playing one game of roulette with a \$2 bet on red or playing two games and betting \$1 on red each time? To find out, we need to compare the probability distributions of the random variables Y = gain from a \$2 bet and T = total gain from two \$1 bets. Which random variable (if either) has the higher expected gain in the long run? Which has the larger variability? By the end of this section, you'll be able to answer questions like these.

Linear Transformations

In Chapter 2, we studied the effects of transformations on the shape, center, and spread of a distribution of data. Recall what we discovered:

1. *Adding (or subtracting) a constant:* Adding the same positive number a to (subtracting a from) each observation:
   - Adds a to (subtracts a from) measures of center and location (mean, median, quartiles, percentiles).
   - Does not change shape or measures of spread (range, *IQR*, standard deviation).
2. Multiplying (or dividing) each observation by the same positive number b:
   - Multiplies (divides) measures of center and location (mean, median, quartiles, percentiles) by b.
   - Multiplies (divides) measures of spread (range, *IQR*, standard deviation) by b.
   - Does not change the shape of the distribution.

How are the probability distributions of random variables affected by similar transformations to the values of the variable? For reasons that will be clear later, we'll start by considering multiplication (or division) by a constant.

Effect of multiplying or dividing by a constant

Let's start with a simple example of a discrete random variable. Pete's Jeep Tours offers a popular half-day trip in a tourist area. There must be at least 2 passengers for the trip to run, and the vehicle will hold up to 6 passengers. The number of passengers X on a randomly selected day has the following probability distribution.

No. of passengers x_i:	2	3	4	5	6
Probability p_i:	0.15	0.25	0.35	0.20	0.05

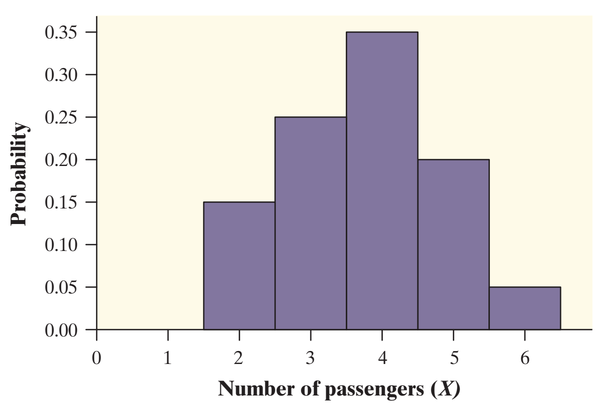


FIGURE 6.5 The probability distribution of the random variable X = the number of passengers on Pete's trip on a randomly chosen day.

Figure 6.5 shows a histogram of the probability distribution.

Using what we learned in Section 6.1, the mean of X is

$$\mu_X = \sum x_i p_i = (2)(0.15) + (3)(0.25) + (4)(0.35) + (5)(0.20) + (6)(0.05) = 3.75$$

That is, Pete expects an average of 3.75 passengers per trip. The variance of X is given by

$$\sigma_X^2 = \sum (x_i - \mu_X)^2 p_i = (2 - 3.75)^2(0.15) + (3 - 3.75)^2(0.25) + \cdots + (6 - 3.75)^2(0.05) = 1.1875$$

So the standard deviation of X is

$$\sigma_X = \sqrt{1.1875} = 1.0897$$

On a randomly selected day, the number of people on a trip typically differs from the mean by about 1.09 passengers.

EXAMPLE Pete's Jeep Tours

Multiplying a random variable by a constant

Pete charges \$150 per passenger. Let C = the total amount of money that Pete collects on a randomly selected trip. Because the amount of money Pete collects is just \$150 times the number of passengers, we can write $C = 150X$. From the probability distribution of X, we can see that the chance of having two people ($X = 2$) on the trip is 0.15. In that case, $C = (150)(2) = 300$. So one possible value of C is \$300, and its corresponding probability is 0.15. If $X = 3$, then $C = (150)(3) = 450$, and the corresponding probability is 0.25. Thus, the probability distribution of C is

Total collected c_i:	300	450	600	750	900
Probability p_i:	0.15	0.25	0.35	0.20	0.05

Figure 6.6 is a histogram of this probability distribution.

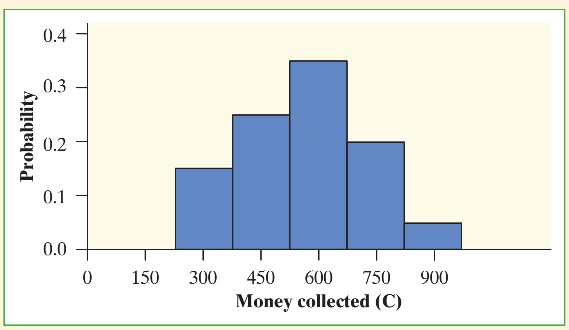


FIGURE 6.6 The probability distribution of the random variable C = the amount of money Pete collects from his trip on a randomly chosen day.

The mean of C is $\mu_C = \sum c_i p_i = (300)(0.15) + (450)(0.25) + \cdots + (900)(0.05) = 562.50$.

On average, Pete will collect a total of \$562.50 from the half-day trip. The variance of C is

$$\begin{aligned}\sigma_C^2 &= \sum(c_i - \mu_C)^2 p_i \\ &= (300 - 562.50)^2(0.15) + (450 - 562.50)^2(0.25) \\ &\quad + \cdots + (900 - 562.50)^2(0.05) = 26{,}718.75\end{aligned}$$

So the standard deviation of C is $\sigma_C = \sqrt{26{,}718.75} = \$163.46$.

In the previous example, the random variable C was obtained by multiplying the values of our earlier random variable X by 150. To understand the effect of multiplying by a constant, let's compare the probability distributions of the random variables X and C.

Shape: The two probability distributions have the same shape.
Center: The mean of X is $\mu_X = 3.75$. The mean of C is $\mu_C = 562.50$, which is (150)(3.75). That is, $\mu_C = 150\mu_X$.
Spread: The standard deviation of X is $\sigma_X = 1.0897$. The standard deviation of C is $\sigma_C = 163.46$, which is (150)(1.0897). That is, $\sigma_C = 150\sigma_X$.

Let's summarize what we've learned so far about transforming a random variable.

EFFECT ON A RANDOM VARIABLE OF MULTIPLYING (OR DIVIDING) BY A CONSTANT

Multiplying (or dividing) each value of a random variable by a positive number b:

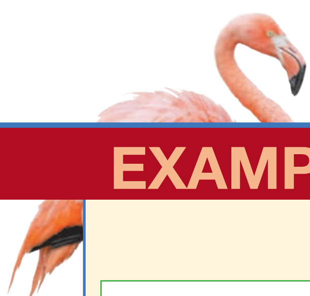
- Multiplies (divides) measures of center and location (mean, median, quartiles, percentiles) by b.
- Multiplies (divides) measures of spread (range, *IQR*, standard deviation) by b.
- Does not change the shape of the distribution.

As with data, if we multiply a random variable by a negative constant b, our common measures of spread are multiplied by $|b|$.

THINK ABOUT IT

How does multiplying by a constant affect the variance? For Pete's Jeep Tours, the variance of the number of passengers on a randomly selected trip is $\sigma_X^2 = 1.1875$. The variance of the total amount of money that Pete collects from such a trip is $\sigma_C^2 = 26{,}718.75$. That's (22,500)(1.1875). So $\sigma_C^2 = 22{,}500\sigma_X^2$. Where did 22,500 come from? It's just $(150)^2$. In other words, $\sigma_C^2 = (150)^2\sigma_X^2$. Multiplying a random variable by a constant b multiplies the variance by b^2.

Effect of adding or subtracting a constant What happens to the probability distribution of a random variable if we add or subtract a constant? Let's return to Pete's Jeep Tours to find out.



EXAMPLE Pete's Jeep Tours

Effect of adding or subtracting a constant

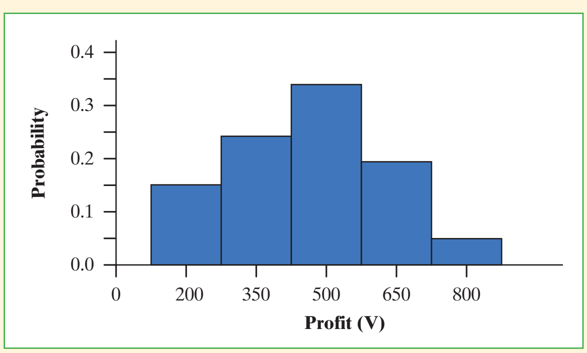


FIGURE 6.7 The probability distribution of the random variable V = the profit that Pete makes from his trip on a randomly chosen day.

It costs Pete \$100 to buy permits, gas, and a ferry pass for each half-day trip. The amount of profit V that Pete makes from the trip is the total amount of money C that he collects from passengers minus \$100. That is, $V = C - 100$. If Pete has only two passengers on the trip ($X = 2$), then $C = 2(150) = 300$ and $V = 200$. From the probability distribution of C, the chance that this happens is 0.15. So the smallest possible value of V is \$200; its corresponding probability is 0.15. If $X = 3$, then $C = 450$ and $V = 350$, and the corresponding probability is 0.25. The probability distribution of V is

Profit v_i:	200	350	500	650	800
Probability p_i:	0.15	0.25	0.35	0.20	0.05

Figure 6.7 shows a histogram of this probability distribution.

The mean of V is $\mu_V = \sum v_i p_i = (200)(0.15) + (350)(0.25) + \cdots + (800)(0.05) = 462.50$. On average, Pete will make a profit of \$462.50 from the trip. The variance of V is

$$\begin{aligned}\sigma_V^2 &= \sum(v_i - \mu_V)^2 p_i \\ &= (200 - 462.50)^2(0.15) + (350 - 462.50)^2(0.25) \\ &\quad + \cdots + (800 - 462.50)^2(0.05) = 26{,}718.75\end{aligned}$$

So the standard deviation of V is

$$\sigma_V = \sqrt{26{,}718.75} = \$163.46$$

It's fairly clear from the previous example that subtracting 100 from the values of the random variable C just shifts the probability distribution to the left by 100. This transformation decreases the mean by 100 (from \$562.50 to \$462.50) but doesn't change the standard deviation (\$163.46) or the shape. These results can be generalized for any random variable.

EFFECT ON A RANDOM VARIABLE OF ADDING (OR SUBTRACTING) A CONSTANT

Adding the same positive number a to (subtracting a from) each value of a random variable:

- Adds a to (subtracts a from) measures of center and location (mean, median, quartiles, percentiles).
- Does not change shape or measures of spread (range, *IQR*, standard deviation).

CHECK YOUR UNDERSTANDING

A large auto dealership keeps track of sales made during each hour of the day. Let X = the number of cars sold during the first hour of business on a randomly selected Friday. Based on previous records, the probability distribution of X is as follows:

Cars sold:	0	1	2	3
Probability:	0.3	0.4	0.2	0.1

The random variable X has mean $\mu_X = 1.1$ and standard deviation $\sigma_X = 0.943$.

1. Suppose the dealership's manager receives a \$500 bonus from the company for each car sold. Let Y = the bonus received from car sales during the first hour on a randomly selected Friday. Find the mean and standard deviation of Y.

2. To encourage customers to buy cars on Friday mornings, the manager spends \$75 to provide coffee and doughnuts. The manager's net profit T on a randomly selected Friday is the bonus earned minus this \$75. Find the mean and standard deviation of T.

Putting it all together: Adding/subtracting and multiplying/dividing What happens if we transform a random variable by both adding or subtracting a constant and multiplying or dividing by a constant? Let's consider

Pete's Jeep Tours again. We could have gone directly from the number of passengers X on a randomly selected jeep tour to Pete's profit V with the equation $V = 150X - 100$ or, equivalently, $V = -100 + 150X$. This **linear transformation** of the random variable X includes both of the transformations that we performed earlier: (1) multiplying by 150 and (2) subtracting 100. (In general, a linear transformation can be written in the form $Y = a + bX$, where a and b are constants.) The net effect of this sequence of transformations is as follows:

For the linear transformation $V = -100 + 150X$, it would *not* be correct to apply the transformations in the reverse order: subtract 100 and then multiply by 150. Doing so would yield the same standard deviation but a different (wrong) mean. Just follow the order of operations from algebra.

Shape: Neither transformation changes the shape of the probability distribution.
Center: The mean of X is multiplied by 150 and then decreased by 100; that is, $\mu_V = 150\mu_X - 100 = -100 + 150\mu_X$.
Spread: The standard deviation of X is multiplied by 150 and is unchanged by the subtraction: $\sigma_V = 150\sigma_X$.

This logic generalizes to any linear transformation.

EFFECTS OF A LINEAR TRANSFORMATION ON A RANDOM VARIABLE

If $Y = a + bX$ is a linear transformation of the random variable X, then

- the probability distribution of Y has the same shape as the probability distribution of X if $b > 0$.
- $\mu_Y = a + b\mu_X$.
- $\sigma_Y = |b|\sigma_X$ (because b could be a negative number).

Can you see why this is called a "linear" transformation? The equation describing the sequence of transformations has the form $Y = a + bX$, which you should recognize as a linear equation.

The bottom two rules in the summary box don't just apply to means and standard deviations. Linear transformations have similar effects on other measures of center or location (median, quartiles, percentiles) and spread (range, IQR). *Whether we're dealing with data or random variables, the effects of a linear transformation are the same.* Note that these results apply to both discrete and continuous random variables.

EXAMPLE The Baby and the Bathwater

Linear transformations

PROBLEM: One brand of bathtub comes with a dial to set the water temperature. When the "babysafe" setting is selected and the tub is filled, the temperature X of the water follows a Normal distribution with a mean of 34°C and a standard deviation of 2°C.

(a) Define the random variable Y to be the water temperature in degrees Fahrenheit (recall that $F = \frac{9}{5}C + 32$) when the dial is set on "babysafe." Find the mean and standard deviation of Y.

(b) According to Babies R Us, the temperature of a baby's bathwater should be between 90°F and 100°F. Find the probability that the water temperature on a randomly selected day when the "babysafe" setting is used meets the Babies R Us recommendation. Show your work.

SOLUTION:

(a) According to the formula for converting Celsius to Fahrenheit, $Y = \frac{9}{5}X + 32$. We could also write this in the form $Y = 32 + \frac{9}{5}X$. The mean of Y is

$$\mu_Y = 32 + \frac{9}{5}\mu_X = 32 + \frac{9}{5}(34) = 93.2°\text{F}$$

The standard deviation of Y is

$$\sigma_Y = \frac{9}{5}\sigma_X = \frac{9}{5}(2) = 3.6°\text{F}$$

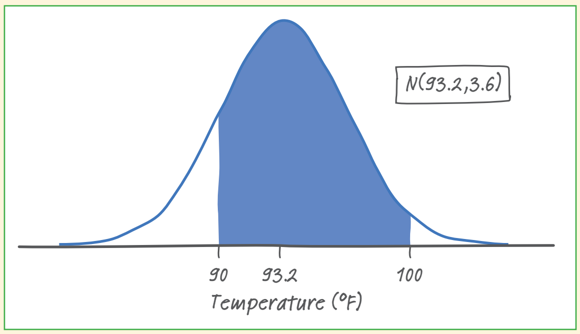


FIGURE 6.8 The Normal probability distribution of the random variable Y = the temperature (in °F) of the bathwater when the dial is set on "babysafe." The shaded area is the probability that the water temperature is between 90°F and 100°F.

(b) **Step 1: State the distribution and the values of interest.** The linear transformation doesn't change the shape of the probability distribution, so the random variable Y is Normally distributed with a mean of 93.2 and a standard deviation of 3.6. We want to find 380$P(90 \le Y \le 100)$. The shaded area in Figure 6.8 shows the desired probability.

Step 2: Perform calculations—show your work! To find this area, we can standardize the boundary values and use Table A:

$$z = \frac{90 - 93.2}{3.6} = -0.89 \quad \text{and} \quad z = \frac{100 - 93.2}{3.6} = 1.89$$

Then $P(-0.89 \le Z \le 1.89) = 0.9706 - 0.1867 = 0.7839$.

Using technology: The command `normalcdf(lower:90, upper:100, μ:93.2, σ:3.6)` gives an area of 0.7835.

Step 3: Answer the question. There's about a 78% chance that the water temperature meets the recommendation on a randomly selected day.

For Practice *Try Exercise* **45**

Combining Random Variables

So far, we have looked at settings that involved a single random variable. Many interesting statistics problems require us to combine two or more random variables.

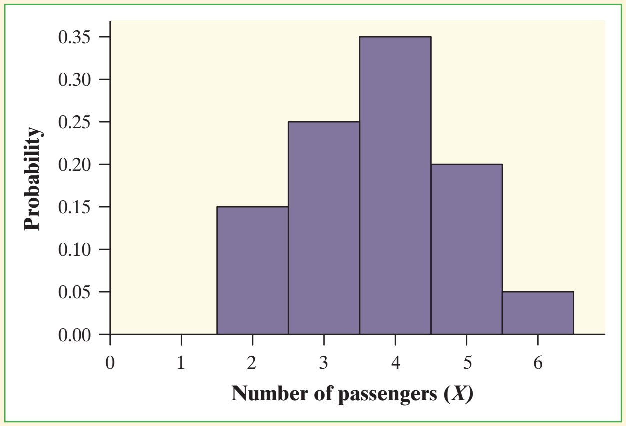
EXAMPLE Pete's Jeeps and Erin's Adventures

When one random variable isn't enough

Earlier, we examined the probability distribution for the random variable X = the number of passengers on a randomly selected half-day trip with Pete's Jeep Tours. Here's a brief recap:

No. of passengers x_i:	2	3	4	5	6
Probability p_i:	0.15	0.25	0.35	0.20	0.05

Mean: $\mu_X = 3.75$ Standard deviation: $\sigma_X = 1.0897$



Pete's sister Erin, who lives near a tourist area in another part of the country, is impressed by the success of Pete's business. She decides to join the business, running tours on the same days as Pete in her slightly smaller vehicle, under the name Erin's Adventures. After a year of steady bookings, Erin discovers that the number of passengers Y on her half-day tours has the following probability distribution. Figure 6.9 displays this distribution as a histogram.

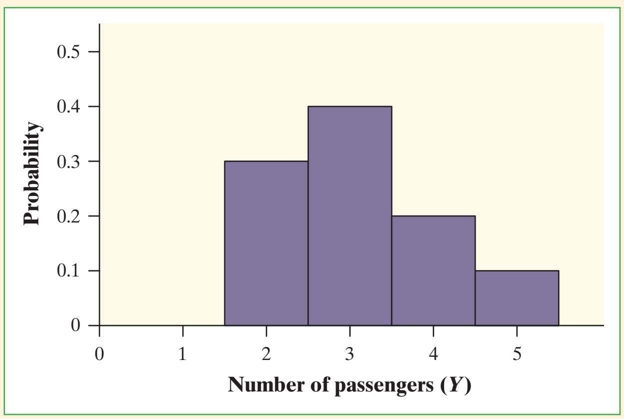


FIGURE 6.9 The probability distribution of the random variable Y = the number of passengers on Erin's trip on a randomly chosen day.

No. of passengers y_i:	2	3	4	5
Probability p_i:	0.3	0.4	0.2	0.1

Mean: $\mu_Y = 3.10$ Standard deviation: $\sigma_Y = 0.943$

How many total passengers T will Pete and Erin have on their tours on a randomly selected day? To answer this question, we need to know about the distribution of the random variable $T = X + Y$.

How many more or fewer passengers D will Pete have than Erin on a randomly selected day? To answer this question, we need to know about the distribution of the random variable $D = X - Y$.

As the example suggests, we want to investigate what happens when we add or subtract random variables.

Sums of random variables

How many total passengers T can Pete and Erin expect to have on their tours on a randomly selected day? Because Pete averages $\mu_X = 3.75$ passengers per trip and Erin averages $\mu_Y = 3.10$ passengers per trip, they will average a total of $\mu_T = 3.75 + 3.10 = 6.85$ passengers per day. We can generalize this result for any two random variables as follows: if $T = X + Y$, then $\mu_T = \mu_X + \mu_Y$. In other words, the expected value (mean) of the sum of two random variables is equal to the sum of their expected values (means).

MEAN OF THE SUM OF RANDOM VARIABLES

For any two random variables X and Y, if $T = X + Y$, then the expected value of T is

$$E(T) = \mu_T = \mu_X + \mu_Y$$

In general, the mean of the sum of several random variables is the sum of their means.

How much variability is there in the total number of passengers who go on Pete's and Erin's tours on a randomly chosen day? Let's think about the possible values of $T = X + Y$. The number of passengers X on Pete's tour is between 2 and 6, and the number of passengers Y on Erin's tour is between 2 and 5. So the total number of passengers T is between 4 and 11. Thus, the range of T is $11 - 4 = 7$. How is this value related to the ranges of X and Y? The range of X is 4 and the range of Y is 3, so

$$\text{range of } T = \text{range of } X + \text{range of } Y$$

That is, there's more variability in the values of T than in the values of X or Y alone. This makes sense, because the variation in X and the variation in Y both contribute to the variation in T.

What about the standard deviation σ_T? If we had the probability distribution of the random variable T, then we could calculate σ_T. Let's try to construct this probability distribution starting with the smallest possible value, $T = 4$. The only way to get a total of 4 passengers is if Pete has $X = 2$ passengers and Erin has $Y = 2$ passengers. We know that $P(X = 2) = 0.15$ and that $P(Y = 2) = 0.3$. If the two events $X = 2$ and $Y = 2$ are independent, then we can multiply these two probabilities. Otherwise, we're stuck. In fact, we can't calculate the probability for any value of T unless X and Y are **independent random variables.**

DEFINITION: Independent random variables

If knowing whether any event involving X alone has occurred tells us nothing about the occurrence of any event involving Y alone, and vice versa, then X and Y are **independent random variables**.

Probability models often assume independence when the random variables describe outcomes that appear unrelated to each other. You should always ask whether the assumption of independence seems reasonable. For instance, it's reasonable to treat the random variables X = number of passengers on Pete's trip and Y = number of passengers on Erin's trip on a randomly chosen day as independent, because the siblings operate their trips in different parts of the country. Now we can calculate the probability distribution of the total number of passengers that day.

EXAMPLE Pete's Jeep Tours and Erin's Adventures

Sum of two random variables

Let $T = X + Y$, as before. Because X and Y are independent random variables, $P(T = 4) = P(X = 2 \text{ and } Y = 2) = P(X = 2) \times P(Y = 2) = (0.15)(0.3) = 0.045$. There are two ways to get a total of $T = 5$ passengers on a randomly selected day: $X = 3, Y = 2$ or $X = 2, Y = 3$. So $P(T = 5) = P(X = 2 \text{ and } Y = 3) + P(X = 3 \text{ and } Y = 2) = (0.15)(0.4) + (0.25)(0.3) = 0.06 + 0.075 = 0.135$.

We can construct the probability distribution by listing all combinations of X and Y that yield each possible value of T and adding the corresponding probabilities. Here is the result.

Value t_i:	4	5	6	7	8	9	10	11
Probability p_i:	0.045	0.135	0.235	0.265	0.190	0.095	0.030	0.005

You can check that the probabilities add to 1. A histogram of the probability distribution is shown in Figure 6.10.

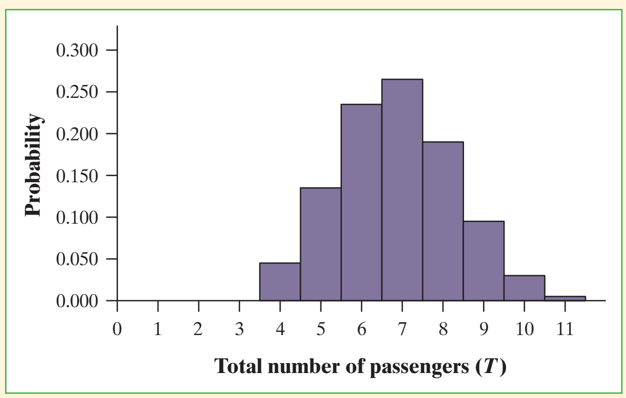


FIGURE 6.10 The probability distribution of the random variable T = the total number of passengers on Pete's and Erin's trips on a randomly chosen day.

The mean of T is $\mu_T = \sum t_i p_i = (4)(0.045) + (5)(0.135) + \cdots + (11)(0.005) = 6.85$. Recall that $\mu_X = 3.75$ and $\mu_Y = 3.10$. Our calculation confirms that

$$\mu_T = \mu_X + \mu_Y = 3.75 + 3.10 = 6.85$$

What about the variance of T? It's

$$\begin{aligned}\sigma_T^2 &= \sum (t_i - \mu_T)^2 p_i \\ &= (4 - 6.85)^2(0.045) + (5 - 6.85)^2(0.135) \\ &\quad + \cdots + (11 - 6.85)^2(0.005) = 2.0775\end{aligned}$$

Recalling that $\sigma_X^2 = 1.1875$ and $\sigma_Y^2 = 0.89$, we see that $1.1875 + 0.89 = 2.0775$. That is,

$$\sigma_T^2 = \sigma_X^2 + \sigma_Y^2$$

To find the standard deviation of T, take the square root of the variance

$$\sigma_T = \sqrt{2.0775} = 1.441$$

As the preceding example illustrates, when we add two *independent* random variables, their variances add. *Standard deviations do not add.* For Pete's and Erin's passenger totals,

$$\sigma_X + \sigma_Y = 1.0897 + 0.943 = 2.0327$$

This is very different from $\sigma_T = 1.441$.

VARIANCE OF THE SUM OF INDEPENDENT RANDOM VARIABLES

For any two *independent* random variables X and Y, if $T = X + Y$, then the variance of T is

$$\sigma_T^2 = \sigma_X^2 + \sigma_Y^2$$

In general, the variance of the sum of several *independent* random variables is the sum of their variances.

You might be wondering whether there's a formula for computing the variance of the sum of two random variables that are *not* independent. There is, but it's beyond the scope of this course. *Just remember that you can add variances only if the two random variables are independent and that you can never add standard deviations.*

caution !

EXAMPLE SAT Scores

The role of independence

A college uses SAT scores as one criterion for admission. Experience has shown that the distribution of SAT scores among its entire population of applicants is such that

SAT Math score X:	$\mu_X = 519$	$\sigma_X = 115$
SAT Critical Reading score Y:	$\mu_Y = 507$	$\sigma_Y = 111$

PROBLEM: What are the mean and standard deviation of the total score $X + Y$ for a randomly selected applicant to this college?

SOLUTION: The mean total score is

$$\mu_{X+Y} = \mu_X + \mu_Y = 519 + 507 = 1026$$

The variance and standard deviation of the total *cannot be computed* from the information given. SAT Math and Critical Reading scores are not independent, because students who score high on one exam tend to score high on the other also.

For Practice *Try Exercise* 47

The next example involves two independent random variables and some transformations.

EXAMPLE Pete's and Erin's Tours

Rules for adding random variables

Earlier, we defined X = the number of passengers that Pete has and Y = the number of passengers that Erin has on a randomly selected day. Recall that

$$\mu_X = 3.75, \sigma_X = 1.0897 \qquad \mu_Y = 3.10, \sigma_Y = 0.943$$

Pete charges \$150 per passenger and Erin charges \$175 per passenger.

PROBLEM: Calculate the mean and the standard deviation of the total amount that Pete and Erin collect on a randomly chosen day.

SOLUTION: Let W = the total amount collected. Then $W = 150X + 175Y$. If we let $C = 150X$ and $G = 175Y$, then we can write W as the sum of two random variables: $W = C + G$. We can use what we learned earlier about the effect of multiplying by a constant to find the mean and standard deviation of C and G. For $C = 150X$,

$$\mu_C = 150\mu_X = 150(3.75) = \$562.50 \text{ and } \sigma_C = 150(1.0897) = \$163.46$$

For $G = 175Y$,

$$\mu_G = 175\mu_Y = 175(3.10) = \$542.50 \text{ and } \sigma_G = 175(0.943) = \$165.03$$

We know that the mean of the sum of two random variables equals the sum of their means:

$$\mu_W = \mu_C + \mu_G = 562.50 + 542.50 = 1105$$

On average, Pete and Erin expect to collect a total of \$1105 per day.

Because the number of passengers X and Y are *independent* random variables, so are the amounts of money collected C and G. Therefore, the variance of W is the sum of the variances of C and G.

$$\sigma_W^2 = \sigma_C^2 + \sigma_G^2 = (163.46)^2 + (165.03)^2 = 53{,}954.07$$

To get the standard deviation, we take the square root of the variance:

$$\sigma_W = \sqrt{53{,}954.07} = 232.28$$

The standard deviation of the total amount they collect is \$232.28.

For Practice *Try Exercise* 51

We can extend our rules for adding random variables to situations involving repeated observations from the same chance process. For instance, suppose a gambler plays two games of roulette, each time placing a \$1 bet on either red or black. What can we say about his total gain (or loss) from playing two games? Earlier, we showed that if X = the amount gained on a single \$1 bet on red or black, then $\mu_X = -\$0.05$ and $\sigma_X = \$1.00$. Because we're interested in the player's total gain over two games, we'll define X_1 as the amount he gains from the first game and X_2 as the amount he gains from the second game. Then, his total gain $T = X_1 + X_2$. Both X_1 and X_2 have the

same probability distribution as X and, therefore, the same mean ($-\$0.05$) and standard deviation ($1.00). The player's expected gain in two games is

$$\mu_T = \mu_{X_1} + \mu_{X_2} = (-\$0.05) + (-\$0.05) = -\$0.10$$

Because knowing the result of one game tells the player nothing about the result of the other game, X_1 and X_2 are independent random variables. As a result,

$$\sigma_T^2 = \sigma_{X_1}^2 + \sigma_{X_2}^2 = (1.00)^2 + (1.00)^2 = 2.00$$

and the standard deviation of the player's total gain is

$$\sigma_T = \sqrt{2.00} = \$1.41$$

THINK ABOUT IT

$X_1 + X_2$ is not the same as $2X$ At the beginning of the section, we asked whether a roulette player would be better off placing two separate $1 bets on red or a single $2 bet on red. The player's total gain from two $1 bets on red is $T = X_1 + X_2$. This sum of random variables has mean $\mu_T = -\$0.10$ and standard deviation $\sigma_T = \$1.41$. Now think about what happens if the gambler places a $2 bet on red in a single game of roulette. Because the random variable X represents a player's gain from a $1 bet, the random variable $Y = 2X$ represents his gain from a $2 bet.

What's the player's expected gain from a single $2 bet on red? It's

$$\mu_Y = 2\mu_X = 2(-\$0.05) = -\$0.10$$

That's the same as his expected gain from playing two games of roulette with a $1 bet each time. But the standard deviation of the player's gain from a single $2 bet is

$$\sigma_Y = 2\sigma_X = 2(\$1.00) = \$2.00$$

Compare this result to $\sigma_T = \$1.41$. There's more variability in the gain from a single $2 bet than in the total gain from two $1 bets.

Let's take this one step further. Would it be better for the player to place a single $100 bet on red or to play 100 games and bet $1 each time on red? For the single $100 bet, the mean and standard deviation of the amount gained would be

$$\begin{aligned} \text{mean} &= 100\mu_X = 100(-\$0.05) = -\$5.00 \\ \text{standard deviation} &= 100\sigma_X = 100(\$1.00) = \$100.00 \end{aligned}$$

For 100 games with a $1 bet, the mean and standard deviation of the amount gained would be

$$\begin{aligned} \text{mean} &= \mu_{X_1} + \mu_{X_2} + \cdots + \mu_{X_{100}} \\ &= (-\$0.05) + (-\$0.05) + \cdots + (-\$0.05) = -\$5.00 \\ \text{variance} &= \sigma_{X_1}^2 + \sigma_{X_2}^2 + \cdots + \sigma_{X_{100}}^2 = (1)^2 + (1)^2 + \cdots + (1)^2 = 100 \\ \text{standard deviation} &= \sqrt{100} = \$10.00 \end{aligned}$$

The player has a much better chance of winning (or losing) big with a single $100 bet than with 100 separate $1 bets. Of course, the casino accepts thousands of bets each day, so it can count on being fairly close to its expected return of 5 cents per dollar bet.

CHECK YOUR UNDERSTANDING

A large auto dealership keeps track of sales and lease agreements made during each hour of the day. Let X = the number of cars sold and Y = the number of cars leased during the first hour of business on a randomly selected Friday. Based on previous records, the probability distributions of X and Y are as follows:

Cars sold x_i:	0	1	2	3
Probability p_i:	0.3	0.4	0.2	0.1

Mean: $\mu_X = 1.1$ Standard deviation: $\sigma_X = 0.943$

Cars leased y_i:	0	1	2
Probability p_i:	0.4	0.5	0.1

Mean: $\mu_Y = 0.7$ Standard deviation: $\sigma_Y = 0.64$

Define $T = X + Y$. Assume that X and Y are independent.

1. Find and interpret μ_T.
2. Compute σ_T. Show your work.
3. The dealership's manager receives a \$500 bonus for each car sold and a \$300 bonus for each car leased. Find the mean and standard deviation of the manager's total bonus B. Show your work.

Differences of random variables Now that we've examined sums of random variables, it's time to investigate the difference of two random variables. Let's start by looking at the difference in the number of passengers that Pete and Erin have on their tours on a randomly selected day, $D = X - Y$. Because Pete averages $\mu_X = 3.75$ passengers per trip and Erin averages $\mu_Y = 3.10$ passengers per trip, the average difference is $\mu_D = 3.75 - 3.10 = 0.65$ passengers. That is, Pete averages 0.65 more passengers per day than Erin does. We can generalize this result for any two random variables as follows: if $D = X - Y$, then $\mu_D = \mu_X - \mu_Y$. In other words, the mean (expected value) of the difference of two random variables is equal to the difference of their means (expected values).

MEAN OF THE DIFFERENCE OF RANDOM VARIABLES

For any two random variables X and Y, if $D = X - Y$, then the mean of D is

$$\mu_D = E(D) = \mu_X - \mu_Y$$

The order of subtraction is important. If we had defined $D = Y - X$, then $\mu_D = \mu_Y - \mu_X = 3.10 - 3.75 = -0.65$. In other words, Erin averages 0.65 fewer passengers than Pete does on a randomly chosen day.

Caution

Earlier, we saw that the variance of the sum of two independent random variables is the sum of their variances. Can you guess what the variance of the *difference* of two independent random variables will be? (If you were thinking something like "the difference of their variances," think again!) Let's return to the jeep tours scenario. On a randomly selected day, the number of passengers X on Pete's tour is between 2 and 6, and the number of passengers Y on Erin's tour is between

2 and 5. So the difference in the number of passengers $D = X - Y$ is between -3 and 4. Thus, the range of D is $4 - (-3) = 7$. How is this value related to the ranges of X and Y? The range of X is 4 and the range of Y is 3, so

$$\text{range of } D = \text{range of } X + \text{range of } Y$$

As with sums of random variables, there's more variability in the values of the difference D than in the values of X or Y alone. This should make sense, because the variation in X and the variation in Y both contribute to the variation in D.

If you follow the process we used earlier with the random variable $T = X + Y$, you can build the probability distribution of $D = X - Y$. Here it is.

Value d_i:	−3	−2	−1	0	1	2	3	4
Probability p_i:	0.015	0.055	0.145	0.235	0.260	0.195	0.080	0.015

You can use the probability distribution to confirm that:

1. $\mu_D = \mu_X - \mu_Y = 3.75 - 3.10 = 0.65$
2. $\sigma_D^2 = \sigma_X^2 + \sigma_Y^2 = 1.1875 + 0.89 = 2.0775$
3. $\sigma_D = \sqrt{2.0775} = 1.441$

Result 2 shows that, just like with addition, when we subtract two independent random variables, variances add.

VARIANCE OF THE DIFFERENCE OF RANDOM VARIABLES

For any two *independent* random variables X and Y, if $D = X - Y$, then the variance of D is

$$\sigma_D^2 = \sigma_X^2 + \sigma_Y^2$$

Let's put our new rules for subtracting random variables to use in a familiar setting.

EXAMPLE Pete's Jeep Tours and Erin's Adventures

Difference of random variables

We have defined several random variables related to Pete's and Erin's tour businesses. For a randomly selected day,

C = amount of money that Pete collects G = amount of money that Erin collects

Here are the means and standard deviations of these random variables:

$$\mu_C = 562.50 \qquad \mu_G = 542.50$$
$$\sigma_C = 163.46 \qquad \sigma_G = 165.03$$

PROBLEM: Calculate the mean and the standard deviation of the difference $D = C - G$ in the amounts that Pete and Erin collect on a randomly chosen day. Interpret each value in context.

SOLUTION: We know that the mean of the difference of two random variables is the difference of their means. That is,

$$\mu_D = \mu_C - \mu_G = 562.50 - 542.50 = 20.00$$

On average, Pete collects \$20 more per day than Erin does. Some days the difference will be more than \$20, other days it will be less, but the average difference after lots of days will be about \$20.

Because the number of passengers X and Y are *independent* random variables, so are the amounts of money collected C and G. Therefore, the variance of D is the sum of the variances of C and G:

$$\sigma_D^2 = \sigma_C^2 + \sigma_G^2 = (163.46)^2 + (165.03)^2 = 53{,}954.07$$
$$\sigma_D = \sqrt{53{,}954.07} = 232.28$$

The value $\sigma_D = \$232.28$ in the example should look familiar. It's the same value we got earlier when we calculated the standard deviation of the total amount that Pete and Erin collect on a randomly chosen day, $\sigma_T = \$232.28$.

The standard deviation of the difference in the amounts collected by Pete and Erin is \$232.28. Even though the *average* difference in the amounts collected is \$20, the difference on individual days will typically vary from the mean by about \$232.

For Practice *Try Exercise* 55

CHECK YOUR UNDERSTANDING

A large auto dealership keeps track of sales and lease agreements made during each hour of the day. Let X = the number of cars sold and Y = the number of cars leased during the first hour of business on a randomly selected Friday. Based on previous records, the probability distributions of X and Y are as follows:

Cars sold x_i:	0	1	2	3
Probability p_i:	0.3	0.4	0.2	0.1

Mean: $\mu_X = 1.1$ Standard deviation: $\sigma_X = 0.943$

Cars leased y_i:	0	1	2
Probability p_i:	0.4	0.5	0.1

Mean: $\mu_Y = 0.7$ Standard deviation: $\sigma_Y = 0.64$

Define $D = X - Y$. Assume that X and Y are independent.

1. Find and interpret μ_D.
2. Compute σ_D. Show your work.
3. The dealership's manager receives a \$500 bonus for each car sold and a \$300 bonus for each car leased. Find the mean and standard deviation of the difference in the manager's bonus for cars sold and leased. Show your work.

Combining Normal Random Variables

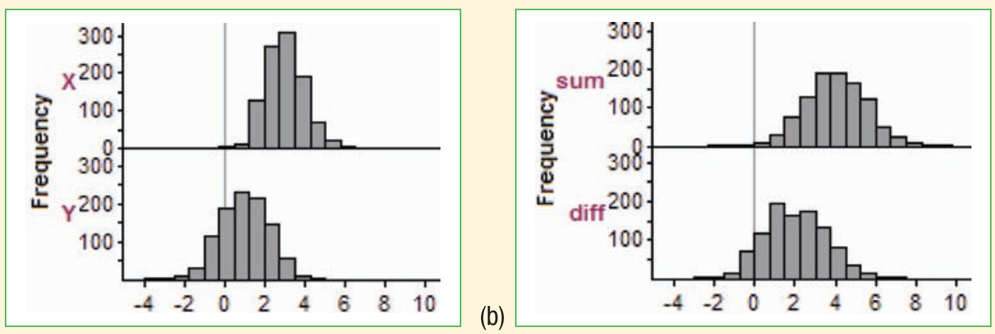
So far, we have concentrated on finding rules for means and variances of random variables. If a random variable is Normally distributed, we can use its mean

and standard deviation to compute probabilities. The earlier example on young women's heights (page 357) shows the method. What if we combine two Normal random variables?

EXAMPLE A Computer Simulation

Sums and differences of Normal random variables

We used Fathom software to simulate taking independent SRSs of 1000 observations from each of two Normally distributed random variables, X and Y. Figure 6.11(a) shows the results. The random variable X is $N(3, 0.9)$ and the random variable Y is $N(1, 1.2)$. What do we know about the sum and difference of these two random variables? The histograms in Figure 6.11(b) came from adding and subtracting the values of X and Y for the 1000 randomly generated observations from each distribution.



(a) (b)

FIGURE 6.11 (a) Histograms showing the results of randomly selecting 1000 values from two different Normal random variables X and Y. (b) Histograms of the sum and difference of the 1000 randomly selected values of X and Y.

Let's summarize what we see:

	Sum X + Y	**Difference X − Y**
Shape:	Looks approximately Normal	Looks approximately Normal
Center:	About 4, which makes sense because $\mu_{X+Y} = \mu_X + \mu_Y = 3+1 = 4$	About 2, which makes sense because $\mu_{X-Y} = \mu_X - \mu_Y = 3 - 1 = 2$

Spread: The spreads of the two distributions are about the same. That makes sense because

$$\sigma^2_{X+Y} = \sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$$

As the previous example illustrates, *any sum or difference of independent Normal random variables is also Normally distributed.* The mean and standard deviation of the resulting Normal distribution can be found using the appropriate rules for means and variances.

EXAMPLE Give Me Some Sugar!

Sums of Normal random variables

Mr. Starnes likes sugar in his hot tea. From experience, he needs between 8.5 and 9 grams of sugar in a cup of tea for the drink to taste right. While making his tea one morning, Mr. Starnes adds four randomly selected packets of sugar. Suppose the amount of sugar in these packets follows a Normal distribution with mean 2.17 grams and standard deviation 0.08 grams.

PROBLEM: What's the probability that Mr. Starnes's tea tastes right?

SOLUTION:

Step 1: State the distribution and the values of interest. Let X = the amount of sugar in a randomly selected packet. Then X_1 = amount of sugar in Packet 1, X_2 = amount of sugar in Packet 2, X_3 = amount of sugar in Packet 3, and X_4 = amount of sugar in Packet 4. Each of these random variables has a Normal distribution with mean 2.17 grams and standard deviation 0.08 grams. We're interested in the total amount of sugar that Mr. Starnes puts in his tea, which is given by $T = X_1 + X_2 + X_3 + X_4$.

The random variable T is a sum of four independent Normal random variables. So T follows a Normal distribution with mean

$$\mu_T = \mu_{X_1} + \mu_{X_2} + \mu_{X_3} + \mu_{X_4} = 2.17 + 2.17 + 2.17 + 2.17 = 8.68 \text{ grams}$$

and variance

$$\sigma_T^2 = \sigma_{X_1}^2 + \sigma_{X_2}^2 + \sigma_{X_3}^2 + \sigma_{X_4}^2 = (0.08)^2 + (0.08)^2 + (0.08)^2 + (0.08)^2 = 0.0256$$

The standard deviation of T is

$$\sigma_T = \sqrt{0.0256} = 0.16 \text{ grams}$$

We want to find the probability that the total amount of sugar in Mr. Starnes's tea is between 8.5 and 9 grams. Figure 6.12 shows this probability as the area under a Normal curve.

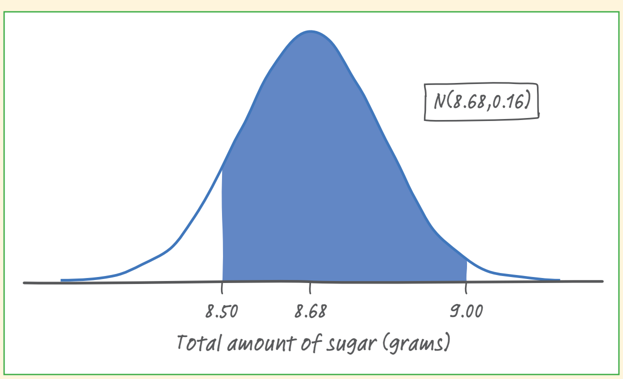


FIGURE 6.12 Normal distribution of the total amount of sugar in Mr. Starnes's tea.

Step 2: Perform calculations—show your work! To find this area, we can standardize the boundary values and use Table A:

$$z = \frac{8.5 - 8.68}{0.16} = -1.13 \quad \text{and} \quad z = \frac{9 - 8.68}{0.16} = 2.00$$

Then $P(-1.13 \le Z \le 2.00) = 0.9772 - 0.1292 = 0.8480$.

Using Technology: The command `normalcdf(lower:8.5, upper:9, μ:8.68, σ:0.16)` gives an area of 0.8470.

Step 3: Answer the question. There's about an 85% chance that Mr. Starnes's tea will taste right.

For Practice *Try Exercise* 61

Here's an example that involves subtracting two Normal random variables.

EXAMPLE Put a Lid on It!

Differences of Normal random variables

The diameter C of a randomly selected large drink cup at a fast-food restaurant follows a Normal distribution with a mean of 3.96 inches and a standard deviation of 0.01 inches. The diameter L of a randomly selected large lid at this restaurant follows a Normal distribution with mean 3.98 inches and standard deviation 0.02 inches. For a lid to fit on a cup, the value of L has to be bigger than the value of C, but not by more than 0.06 inches.

PROBLEM: What's the probability that a randomly selected large lid will fit on a randomly chosen large drink cup?

SOLUTION:

Step 1: State the distribution and the values of interest. We'll define the random variable $D = L - C$ to represent the difference between the lid's diameter and the cup's diameter.

The random variable D is the difference of two independent Normal random variables. So D follows a Normal distribution with mean

$$\mu_D = \mu_L - \mu_C = 3.98 - 3.96 = 0.02$$

and variance

$$\sigma_D^2 = \sigma_L^2 + \sigma_C^2 = (0.02)^2 + (0.01)^2 = 0.0005$$

The standard deviation of D is

$$\sigma_D = \sqrt{0.0005} = 0.0224$$

We want to find the probability that the difference D is between 0 and 0.06 inches. Figure 6.13 shows this probability as the area under a Normal curve.

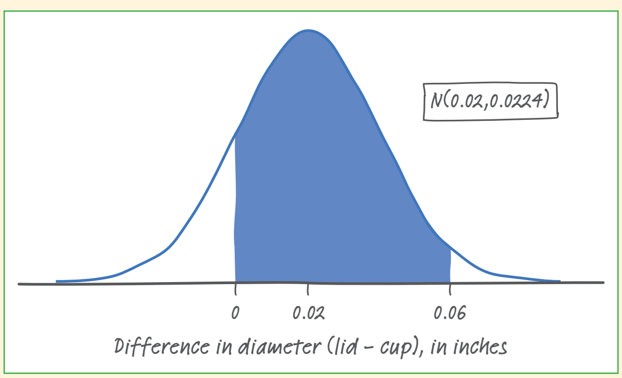


FIGURE 6.13 Normal distribution of the difference in lid diameter and cup diameter at a fast-food restaurant.

Step 2: Perform calculations—show your work! To find this area, we can standardize the boundary values and use Table A:

$$z = \frac{0 - 0.02}{0.0224} = -0.89 \quad \text{and} \quad z = \frac{0.06 - 0.02}{0.0224} = 1.79$$

Then $P(-0.89 \leq Z \leq 1.79) = 0.9633 - 0.1867 = 0.7766$.

Using Technology: The command `normalcdf(lower:0, upper:0.06, μ:0.02, σ:0.0224)` gives an area of 0.7770.

Step 3: Answer the question. There's about a 78% chance that a randomly selected large lid will fit on a randomly chosen large drink cup at this fast-food restaurant. Roughly 22% of the time, the lid won't fit. This seems like an unreasonably high chance of getting a lid that doesn't fit. Maybe the restaurant should find a new supplier!

For Practice *Try Exercise* 63

Section 6.2 Summary

- Adding a positive constant a to (subtracting a from) a random variable increases (decreases) the mean of the random variable by a but does not affect its standard deviation or the shape of its probability distribution.
- Multiplying (dividing) a random variable by a positive constant b multiplies (divides) the mean of the random variable by b and the standard deviation by b but does not change the shape of its probability distribution.
- A **linear transformation** of a random variable involves adding or subtracting a constant a, multiplying or dividing by a constant b, or both. We can write a linear transformation of the random variable X in the form $Y = a + bX$. The shape, center, and spread of the probability distribution of Y are as follows:

  Shape: Same as the probability distribution of X if $b > 0$.

  Center: $\mu_Y = a + b\mu_X$

  Spread: $\sigma_Y = |b|\sigma_X$
- If X and Y are *any* two random variables,

  $\mu_{X+Y} = \mu_X + \mu_Y$: The mean of the sum of two random variables is the sum of their means.

  $\mu_{X-Y} = \mu_X - \mu_Y$: The mean of the difference of two random variables is the difference of their means.
- If X and Y are **independent random variables**, then knowing the value of one variable tells you nothing about the value of the other. In that case, variances add:

  $\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$: The variance of the sum of two independent random variables is the sum of their variances.

  $\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$: The variance of the difference of two independent random variables is the sum of their variances.
- The sum or difference of independent Normal random variables follows a Normal distribution.

Section 6.2 Exercises

35. **Crickets** The length in inches of a cricket chosen at random from a field is a random variable X with mean 1.2 inches and standard deviation 0.25 inches. Find the mean and standard deviation of the length Y of a randomly chosen cricket from the field in centimeters. There are 2.54 centimeters in an inch.

36. **Men's heights** A report of the National Center for Health Statistics says that the height of a 20-year-old man chosen at random is a random variable H with mean 5.8 feet and standard deviation 0.24 feet. Find the mean and standard deviation of the height J of a randomly selected 20-year-old man in inches. There are 12 inches in a foot.

37. **Get on the boat!** A small ferry runs every half hour from one side of a large river to the other. The number of cars X on a randomly chosen ferry trip has the probability distribution shown below. You can check that $\mu_X = 3.87$ and $\sigma_X = 1.29$.

Cars:	0	1	2	3	4	5
Probability:	0.02	0.05	0.08	0.16	0.27	0.42

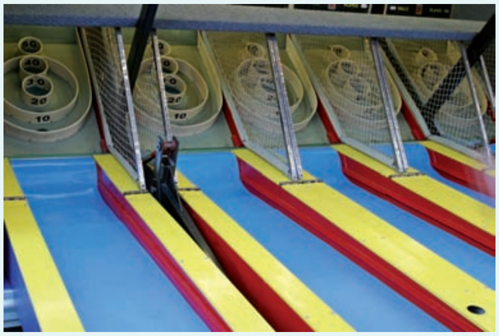
(a) The cost for the ferry trip is \$5. Make a graph of the probability distribution for the random variable M = money collected on a randomly selected ferry trip. Describe its shape.

(b) Find and interpret μ_M.

(c) Find and interpret σ_M.

38. **Skee Ball** Ana is a dedicated Skee Ball player (see photo) who always rolls for the 50-point slot. The probability distribution of Ana's score X on a single roll of the ball is shown below. You can check that $\mu_X = 23.8$ and $\sigma_X = 12.63$.

Score:	10	20	30	40	50
Probability:	0.32	0.27	0.19	0.15	0.07



(a) A player receives one ticket from the game for every 10 points scored. Make a graph of the probability distribution for the random variable T = number of tickets Ana gets on a randomly selected throw. Describe its shape.

(b) Find and interpret μ_T.

(c) Find and interpret σ_T.

Exercises 39 and 40 refer to the following setting. Ms. Hall gave her class a 10-question multiple-choice quiz. Let X = the number of questions that a randomly selected student in the class answered correctly. The computer output below gives information about the probability distribution of X. To determine each student's grade on the quiz (out of 100), Ms. Hall will multiply his or her number of correct answers by 5 and then add 50. Let G = the grade of a randomly chosen student in the class.

N	Mean	Median	StDev	Min	Max	Q_1	Q_3
30	7.6	8.5	1.32	4	10	8	9

39. **Easy quiz**

(a) Find the mean of G. Show your method.

(b) Find the standard deviation of G. Show your method.

(c) How do the variance of X and the variance of G compare? Justify your answer.

40. **Easy quiz**

(a) Find the median of G. Show your method.

(b) Find the *IQR* of G. Show your method.

(c) What shape would the probability distribution of G have? Justify your answer.

41. **Get on the boat!** Refer to Exercise 37. The ferry company's expenses are \$20 per trip. Define the random variable Y to be the amount of profit (money collected minus expenses) made by the ferry company on a randomly selected trip. That is, $Y = M - 20$.

(a) Find and interpret the mean of Y.

(b) Find and interpret the standard deviation of Y.

42. **The Tri-State Pick 3** Most states and Canadian provinces have government-sponsored lotteries. Here is a simple lottery wager, from the Tri-State Pick 3 game that New Hampshire shares with Maine and Vermont. You choose a number with 3 digits from 0 to 9; the state chooses a three-digit winning number at random and pays you \$500 if your number is chosen. Because there are 1000 numbers with three digits, you have probability 1/1000 of winning. Taking X to be the amount your ticket pays you, the probability distribution of X is

Payoff:	\$0	\$500
Probability:	0.999	0.001

(a) Show that the mean and standard deviation of X are $\mu_X = \$0.50$ and $\sigma_X = \$15.80$.

(b) If you buy a Pick 3 ticket, your winnings are $W = X - 1$, because it costs \$1 to play. Find the mean and standard deviation of W. Interpret each of these values in context.

43. **Get on the boat!** Based on the analysis in Exercise 41, the ferry company decides to increase the cost of a trip to \$6. We can calculate the company's profit Y on a randomly selected trip from the number of cars X. Find the mean and standard deviation of Y. Show your work.

44. **Making a profit** Rotter Partners is planning a major investment. From experience, the amount of profit X (in millions of dollars) on a randomly selected investment of this type is uncertain, but an estimate gives the following probability distribution:

Profit:	1	1.5	2	4	10
Probability:	0.1	0.2	0.4	0.2	0.1

Based on this estimate, $\mu_X = 3$ and $\sigma_X = 2.52$. Rotter Partners owes its lender a fee of \$200,000 plus 10% of the profits X. So the firm actually retains $Y = 0.9X - 0.2$ from the investment. Find the mean and standard deviation of Y. Show your work.

45. **Too cool at the cabin?** During the winter months, the temperatures at the Starneses' Colorado cabin can stay well below freezing (32°F or 0°C) for weeks at a time. To prevent the pipes from freezing, Mrs. Starnes sets the thermostat at 50°F. She also buys a digital thermometer that records the indoor temperature each night at midnight. Unfortunately, the thermometer is programmed to measure the temperature in degrees Celsius. Based on several years' worth of data, the temperature T in the cabin at midnight on a randomly selected night follows a Normal distribution with mean 8.5°C and standard deviation 2.25°C.

pg 368

(a) Let Y = the temperature in the cabin at midnight on a randomly selected night in degrees Fahrenheit (recall that $F = (9/5)C + 32$). Find the mean and standard deviation of Y.

(b) Find the probability that the midnight temperature in the cabin is below 40°F. Show your work.

46. **Cereal** A company's single-serving cereal boxes advertise 9.63 ounces of cereal. In fact, the amount of cereal X in a randomly selected box follows a Normal distribution with a mean of 9.70 ounces and a standard deviation of 0.03 ounces.

(a) Let Y = the excess amount of cereal beyond what's advertised in a randomly selected box, measured in grams (1 ounce = 28.35 grams). Find the mean and standard deviation of Y.

(b) Find the probability of getting at least 3 grams more cereal than advertised. Show your work.

47. **His and her earnings** Researchers randomly select a married couple in which both spouses are employed. Let X be the income of the husband and Y be the income of the wife. Suppose that you know the means μ_X and μ_Y and the variances σ_X^2 and σ_Y^2 of both variables.

pg 373

(a) Is it reasonable to take the mean of the total income $X + Y$ to be $\mu_X + \mu_Y$? Explain your answer.

(b) Is it reasonable to take the variance of the total income to be $\sigma_X^2 + \sigma_Y^2$? Explain your answer.

48. **Rainy days** Imagine that we randomly select a day from the past 10 years. Let X be the recorded rainfall on this date at the airport in Orlando, Florida, and Y be the recorded rainfall on this date at Disney World just outside Orlando. Suppose that you know the means μ_X and μ_Y and the variances σ_X^2 and σ_Y^2 of both variables.

(a) Is it reasonable to take the mean of the total rainfall $X + Y$ to be $\mu_X + \mu_Y$? Explain your answer.

(b) Is it reasonable to take the variance of the total rainfall to be $\sigma_X^2 + \sigma_Y^2$? Explain your answer.

49. **Get on the boat!** Refer to Exercise 41. Find the expected value and standard deviation of the total amount of profit made on two randomly selected days. Show your work.

50. **The Tri-State Pick 3** Refer to Exercise 42. Suppose you buy one Pick 3 ticket on each of two consecutive days. Find the expected value and standard deviation of your total winnings. Show your work.

51. **Essay errors** Typographical and spelling errors can be either "nonword errors" or "word errors." A nonword error is not a real word, as when "the" is typed as "teh." A word error is a real word, but not the right word, as when "lose" is typed as "loose." When students are asked to write a 250-word essay (without spell-checking), the number of nonword errors X in a randomly selected essay has the following probability distribution:

pg 374

Value:	0	1	2	3	4
Probability:	0.1	0.2	0.3	0.3	0.1

$$\mu_X = 2.1 \qquad \sigma_X = 1.136$$

The number of word errors Y has this probability distribution:

Value:	0	1	2	3
Probability:	0.4	0.3	0.2	0.1

$$\mu_Y = 1.0 \qquad \sigma_Y = 1.0$$

Assume that X and Y are independent.

An English professor deducts 3 points from a student's essay score for each nonword error and 2 points for each word error. Find the mean and standard deviation of the total score deductions for a randomly selected essay. Show your work.

52. **The Tri-State Pick 3** Refer to Exercise 42. You and a friend decide to play Pick 3, but with two different strategies. Your friend buys a \$1 Pick 3 ticket on each of five consecutive days. You bet \$5 on a single number on your Pick 3 ticket. Find the mean and standard deviation of the total winnings for you and your friend. Show your work.

53. **Essay errors** Refer to Exercise 51.

(a) Find the mean and standard deviation of the difference $Y - X$ in the number of errors made by a randomly selected student. Interpret each value in context.

(b) From the information given, can you find the probability that a randomly selected student makes more word errors than nonword errors? If so, find this probability. If not, explain why not.

54. **Study habits** The Survey of Study Habits and Attitudes (SSHA) is a psychological test that measures academic motivation and study habits. The distribution of SSHA scores among the women at a college has mean 120 and standard deviation 28, and the distribution of scores among male students has mean

105 and standard deviation 35. You select a single male student and a single female student at random and give them the SSHA test.

(a) Find the mean and standard deviation of the difference (female minus male) between their scores. Interpret each value in context.

(b) From the information given, can you find the probability that the woman chosen scores higher than the man? If so, find this probability. If not, explain why you cannot.

55. **Essay scores** Refer to Exercise 51. Find the mean and standard deviation of the difference in score deductions (nonword – word) for a randomly selected essay. Show your work. pg 377

56. **The Tri-State Pick 3** Refer to Exercise 52. Find the mean and standard deviation of the difference (you – your friend) in winnings. Show your work.

Exercises 57 and 58 refer to the following setting. In Exercises 14 and 18 of Section 6.1, we examined the probability distribution of the random variable X = the amount a life insurance company earns on a randomly chosen 5-year term life policy. Calculations reveal that $\mu_X = \$303.35$ and $\sigma_X = \$9707.57$.

57. **Life insurance** The risk of insuring one person's life is reduced if we insure many people. Suppose that we insure two 21-year-old males, and that their ages at death are independent. If X_1 and X_2 are the insurer's income from the two insurance policies, the insurer's average income W on the two policies is

$$W = \frac{X_1 + X_2}{2}$$

Find the mean and standard deviation of W. (You see that the mean income is the same as for a single policy but the standard deviation is less.)

58. **Life insurance** If four 21-year-old men are insured, the insurer's average income is

$$V = \frac{X_1 + X_2 + X_3 + X_4}{4}$$

where X_i is the income from insuring one man. Assuming that the amount of income earned on individual policies is independent, find the mean and standard deviation of V. (If you compare with the results of Exercise 57, you should see that averaging over more insured individuals reduces risk.)

59. **Time and motion** A time-and-motion study measures the time required for an assembly-line worker to perform a repetitive task. The data show that the time required to bring a part from a bin to its position on an automobile chassis varies from car to car according to a Normal distribution with mean 11 seconds and standard deviation 2 seconds. The time required to attach the part to the chassis follows a Normal distribution with mean 20 seconds and standard deviation 4 seconds. The study finds that the times required for the two steps are independent. A part that takes a long time to position, for example, does not take more or less time to attach than other parts.

(a) What is the distribution of the time required for the entire operation of positioning and attaching a randomly selected part?

(b) Management's goal is for the entire process to take less than 30 seconds. Find the probability that this goal will be met for a randomly selected part.

60. **Electronic circuit** The design of an electronic circuit for a toaster calls for a 100-ohm resistor and a 250-ohm resistor connected in series so that their resistances add. The components used are not perfectly uniform, so that the actual resistances vary independently according to Normal distributions. The resistance of 100-ohm resistors has mean 100 ohms and standard deviation 2.5 ohms, while that of 250-ohm resistors has mean 250 ohms and standard deviation 2.8 ohms.

(a) What is the distribution of the total resistance of the two components in series for a randomly selected toaster?

(b) Find the probability that the total resistance for a randomly selected toaster lies between 345 and 355 ohms.

61. **Swim team** Hanover High School has the best women's swimming team in the region. The 400-meter freestyle relay team is undefeated this year. In the 400-meter freestyle relay, each swimmer swims 100 meters. The times, in seconds, for the four swimmers this season are approximately Normally distributed with means and standard deviations as shown. Assume that the swimmer's individual times are independent. Find the probability that the total team time in the 400-meter freestyle relay for a randomly selected race is less than 220 seconds. pg 380

Swimmer	Mean	Std. dev.
Wendy	55.2	2.8
Jill	58.0	3.0
Carmen	56.3	2.6
Latrice	54.7	2.7

62. **Toothpaste** Ken is traveling for his business. He has a new 0.85-ounce tube of toothpaste that's supposed to last him the whole trip. The amount of toothpaste Ken squeezes out of the tube each time he brushes varies according to a Normal distribution with mean 0.13 ounces and standard deviation 0.02 ounces. If Ken brushes his teeth six times on a randomly selected trip, what's the probability that he'll use all the toothpaste in the tube?

63. **Auto emissions** The amount of nitrogen oxides (NOX) present in the exhaust of a particular type of car varies from car to car according to a Normal distribution with mean 1.4 grams per mile (g/mi) and standard deviation 0.3 g/mi. Two randomly selected cars of this type are tested. One has 1.1 g/mi of NOX; the other has 1.9 g/mi. The test station attendant finds this difference in emissions between two similar cars surprising. If the NOX levels for two randomly chosen cars of this type are independent, find the probability that the difference is at least as large as the value the attendant observed.

pg 381

64. **Loser buys the pizza** Leona and Fred are friendly competitors in high school. Both are about to take the ACT college entrance examination. They agree that if one of them scores 5 or more points better than the other, the loser will buy the winner a pizza. Suppose that in fact Fred and Leona have equal ability, so that each score varies Normally with mean 24 and standard deviation 2. (The variation is due to luck in guessing and the accident of the specific questions being familiar to the student.) The two scores are independent. What is the probability that the scores differ by 5 or more points in either direction?

Multiple choice: Select the best answer for Exercises 65 and 66, which refer to the following setting. The number of calories in a 1-ounce serving of a certain breakfast cereal is a random variable with mean 110 and standard deviation 10. The number of calories in a cup of whole milk is a random variable with mean 140 and standard deviation 12. For breakfast, you eat 1 ounce of the cereal with 1/2 cup of whole milk. Let T be the random variable that represents the total number of calories in this breakfast.

65. The mean of T is

(a) 110. (b) 140. (c) 180. (d) 195. (e) 250.

66. The standard deviation of T is

(a) 22. (b) 16. (c) 15.62. (d) 11.66. (e) 4.

67. **Statistics for investing** (3.1) Joe's retirement plan invests in stocks through an "index fund" that follows the behavior of the stock market as a whole, as measured by the Standard & Poor's (S&P) 500 stock index. Joe wants to buy a mutual fund that does not track the index closely. He reads that monthly returns from Fidelity Technology Fund have correlation $r = 0.77$ with the S&P 500 index and that Fidelity Real Estate Fund has correlation $r = 0.37$ with the index.

(a) Which of these funds has the closer relationship to returns from the stock market as a whole? How do you know?

(b) Does the information given tell Joe anything about which fund has had higher returns?

68. **Buying stock** (5.3, 6.1) You purchase a hot stock for \$1000. The stock either gains 30% or loses 25% each day, each with probability 0.5. Its returns on consecutive days are independent of each other. You plan to sell the stock after two days.

(a) What are the possible values of the stock after two days, and what is the probability for each value? What is the probability that the stock is worth more after two days than the \$1000 you paid for it?

(b) What is the mean value of the stock after two days? (*Comment:* You see that these two criteria give different answers to the question "Should I invest?")

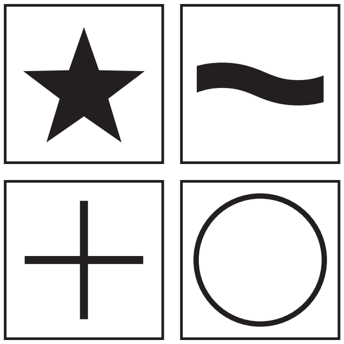
6.3 Binomial and Geometric Random Variables

WHAT YOU WILL LEARN **By the end of the section, you should be able to:**

- Determine whether the conditions for using a binomial random variable are met.
- Compute and interpret probabilities involving binomial distributions.
- Calculate the mean and standard deviation of a binomial random variable. Interpret these values in context.
- Find probabilities involving geometric random variables.
- When appropriate, use the Normal approximation to the binomial distribution to calculate probabilities.*

*This topic is not required for the AP® Statistics exam. Some teachers prefer to discuss this topic when presenting the sampling distribution of $\hat{p}$ (Chapter 7).

When the same chance process is repeated several times, we are often interested in whether a particular outcome does or doesn't happen on each repetition. Here are some examples:



- To test whether someone has extrasensory perception (ESP), choose one of four cards at random—a star, wave, cross, or circle. Ask the person to identify the card without seeing it. Do this a total of 50 times and see how many cards the person identifies correctly. *Chance process:* choose a card at random. *Outcome of interest:* person identifies card correctly. *Random variable:* X = number of correct identifications.
- A shipping company claims that 90% of its shipments arrive on time. To test this claim, take a random sample of 100 shipments made by the company last month and see how many arrived on time. *Chance process:* Randomly select a shipment and check when it arrived. *Outcome of interest:* arrived on time. *Random variable:* Y = number of on-time shipments.
- In the game of Pass the Pigs, a player rolls a pair of pig-shaped dice. On each roll, the player earns points according to how the pigs land. If the player gets a "pig out," in which the two pigs land on opposite sides, she loses all points earned in that round and must pass the pigs to the next player. A player can choose to stop rolling at any point during her turn and to keep the points that she has earned before passing the pigs. *Chance process:* roll the pig dice. *Outcome of interest:* pig out. *Random variable:* T = number of rolls until the player pigs out.

Some random variables, like X and Y in the first two examples above, count the number of times the outcome of interest occurs in a fixed number of repetitions. They are called *binomial random variables*. Other random variables, like T in the Pass the Pigs setting, count the number of repetitions of the chance process it takes for the outcome of interest to occur. They are known as *geometric random variables*. These two special types of discrete random variables are the focus of this section.

Binomial Settings and Binomial Random Variables

What do the following scenarios have in common?

- Toss a coin 5 times. Count the number of heads.
- Spin a roulette wheel 8 times. Record how many times the ball lands in a red slot.
- Take a random sample of 100 babies born in U.S. hospitals today. Count the number of females.

In each case, we're performing repeated *trials* of the same chance process. The number of trials is fixed in advance. Also, knowing the outcome of one trial tells us nothing about the outcome of any other trial. That is, the trials are *independent*. We're interested in the number of times that a specific event (we'll call it a "success") occurs. Our chances of getting a "success" are the same on each trial. When these conditions are met, we have a **binomial setting.**

DEFINITION: Binomial setting

A **binomial setting** arises when we perform several independent trials of the same chance process and record the number of times that a particular outcome occurs. The four conditions for a binomial setting are

- **B**inary? The possible outcomes of each trial can be classified as "success" or "failure."
- **I**ndependent? Trials must be independent; that is, knowing the result of one trial must not tell us anything about the result of any other trial.
- **N**umber? The number of trials n of the chance process must be fixed in advance.
- **S**uccess? There is the same probability p of success on each trial.

The boldface letters in the box give you a helpful way to remember the conditions for a binomial setting: just check the BINS!

When checking the Binary condition, note that there can be more than two possible outcomes per trial—a roulette wheel has numbered slots of *three* colors: red, black, and green. If we define "success" as having the ball land in a red slot, then "failure" occurs when the ball lands in a black or a green slot.

Think of tossing a coin n times as an example of the binomial setting. Each toss gives either heads or tails. Knowing the outcome of one toss doesn't change the probability of a head on any other toss, so the tosses are independent. If we call heads a success, then p is the probability of a head and remains the same as long as we toss the same coin. For tossing a fair coin, p is 0.5. The number of heads we count is a **binomial random variable** X. The probability distribution of X is called a **binomial distribution.**

DEFINITION: Binomial random variable and binomial distribution

The count X of successes in a binomial setting is a **binomial random variable.** The probability distribution of X is a **binomial distribution** with parameters n and p, where n is the number of trials of the chance process and p is the probability of a success on any one trial. The possible values of X are the whole numbers from 0 to n.

Later in the section, we'll learn how to assign probabilities to outcomes and how to find the mean and standard deviation of a binomial random variable. For now, it's important to be able to distinguish situations in which a binomial distribution does and doesn't apply.

EXAMPLE From Blood Types to Aces

Binomial settings and random variables

PROBLEM: *Here are three scenarios involving chance behavior. In each case, determine whether or not the given random variable has a binomial distribution. Justify your answer.*

(a) *Genetics says that children receive genes from each of their parents independently. Each child of a particular set of parents has probability 0.25 of having type O blood. Suppose these parents have 5 children. Let X = the number of children with type O blood.*

(b) Shuffle a deck of cards. Turn over the first 10 cards, one at a time. Let Y = the number of aces you observe.

(c) Shuffle a deck of cards. Turn over the top card. Put the card back in the deck, and shuffle again. Repeat this process until you get an ace. Let W = the number of cards required.

SOLUTION:

(a) To see if this is a binomial setting, we'll check the BINS:

- Binary? "Success" = has type O blood. "Failure" = doesn't have type O blood.
- Independent? Knowing one child's blood type tells you nothing about another's because children inherit genes determining blood type independently from each of their parents.
- Number? There are $n = 5$ trials of this chance process.
- Success? The probability of a success is $p = 0.25$ on each trial.

This is a binomial setting. Because X counts the number of successes, it is a binomial random variable with parameters $n = 5$ and $p = 0.25$.

(b) Let's check the BINS:

- Binary? "Success" = get an ace. "Failure" = don't get an ace.
- Independent? No. If the first card you turn over is an ace, then the next card is less likely to be an ace because you're not replacing the top card in the deck. Similarly, if the first card isn't an ace, the second card is more likely to be an ace.
- Number? There are $n = 10$ trials of this chance process.
- Success? The probability that any particular card in a shuffled deck is an ace is $p = 4/52$.

Because the trials are not independent, this is not a binomial setting.

The Independent condition involves *conditional* probabilities. P(2nd card ace | 1st card ace) = 3/51 ≠ P(2nd card ace) = 4/52, so the trials are not independent. The Success condition is about *unconditional* probabilities. P(kth card in a shuffled deck is an ace) = 4/52, so this condition is met.

(c) Let's check the BINS:

- Binary? "Success" = get an ace. "Failure" = don't get an ace.
- Independent? Because you are replacing the card in the deck and shuffling each time, the result of one trial does not tell you anything about the outcome of any other trial.
- Number? The number of trials is not set in advance. You could get an ace on the first card you turn over, or it may take many cards to get an ace.
- Success? The probability of getting an ace is $p = 4/52$ on each trial.

Because there is no fixed number of trials, this is not a binomial setting.

For Practice *Try Exercises* 69 *and* 71

Part (c) of the example raises an important point about binomial random variables. Besides checking the BINS, make sure that you're being asked to count the number of successes in a certain number of trials. In part (c), you're asked to count the number of *trials* until you get a success. That can't be a binomial random variable. (As you'll see later, W is a *geometric* random variable.)

CHECK YOUR UNDERSTANDING

For each of the following situations, determine whether the given random variable has a binomial distribution or not. Justify your answer.

1. Shuffle a deck of cards. Turn over the top card. Put the card back in the deck, and shuffle again. Repeat this process 10 times. Let X = the number of aces you observe.

2. Choose students at random from your class. Let Y = the number who are over 6 feet tall.

3. Roll a fair die 100 times. Sometime during the 100 rolls, one corner of the die chips off. Let W = the number of 5s you roll.

Binomial Probabilities

In a binomial setting, we can define a random variable (say, X) as the number of successes in n independent trials. What's the probability distribution of X? Let's see if an example can help shed some light on this question.

EXAMPLE

Inheriting Blood Type

Calculating binomial probabilities

Each child of a particular set of parents has probability 0.25 of having type O blood. Genetics says that children receive genes from each of their parents independently. If these parents have 5 children, the count X of children with type O blood is a binomial random variable with $n = 5$ trials and probability $p = 0.25$ of success on each trial. In this setting, a child with type O blood is a "success" (S) and a child with another blood type is a "failure" (F).

What's $P(X = 0)$? That is, what's the probability that *none* of the 5 children has type O blood? It's the chance that all 5 children *don't* have type O blood. The probability that any one of this couple's children doesn't have type O blood is $1 - 0.25 = 0.75$ (complement rule). By the multiplication rule for independent events (Chapter 5),

$$P(X = 0) = P(\text{FFFFF}) = (0.75)(0.75)(0.75)(0.75)(0.75) = (0.75)^5 = 0.2373$$

How about $P(X = 1)$? There are several different ways in which exactly 1 of the 5 children could have type O blood. For instance, the first child born might have type O blood, while the remaining 4 children don't have type O blood. The probability that this happens is

$$P(\text{SFFFF}) = (0.25)(0.75)(0.75)(0.75)(0.75) = (0.25)(0.75)^4$$

Alternatively, Child 2 could be the one that has type O blood. The corresponding probability is

$$P(\text{FSFFF}) = (0.75)(0.25)(0.75)(0.75)(0.75) = (0.25)(0.75)^4$$

There are three more possibilities to consider:

$$P(\text{FFSFF}) = (0.75)(0.75)(0.25)(0.75)(0.75) = (0.25)(0.75)^4$$
$$P(\text{FFFSF}) = (0.75)(0.75)(0.75)(0.25)(0.75) = (0.25)(0.75)^4$$
$$P(\text{FFFFS}) = (0.75)(0.75)(0.75)(0.75)(0.25) = (0.25)(0.75)^4$$

In all, there are five different ways in which exactly 1 child would have type O blood, each with the same probability of occurring. As a result,

$$
\begin{aligned}
P(X = 1) &= P(\text{exactly 1 child with type O blood}) \\
&= P(\text{SFFFF}) + P(\text{FSFFF}) + P(\text{FFSFF}) + P(\text{FFFSF}) + P(\text{FFFFS}) \\
&= 5(0.25)(0.75)^4 = 0.39551
\end{aligned}
$$

There's about a 40% chance that exactly 1 of the couple's 5 children will have type O blood.

Let's continue with the scenario from the previous example. What if we wanted to find $P(X = 2)$, the probability that exactly 2 of the couple's children have type O blood? Because the method doesn't depend on the specific setting, we use "S" for success and "F" for failure for short.

Do the work in two stages, as shown in the example.

- Find the probability that a specific 2 of the 5 tries—say, the first and the third—give successes. This is the outcome SFSFF. Because tries are independent, the multiplication rule for independent events applies. The probability we want is

$$
\begin{aligned}
P(\text{SFSFF}) &= P(\text{S})P(\text{F})P(\text{S})P(\text{F})P(\text{F}) \\
&= (0.25)(0.75)(0.25)(0.75)(0.75) \\
&= (0.25)^2(0.75)^3
\end{aligned}
$$

- Observe that *any one arrangement* of 2 S's and 3 F's has this same probability. This is true because we multiply together 0.25 twice and 0.75 three times whenever we have 2 S's and 3 F's. The probability that $X = 2$ is the probability of getting 2 S's and 3 F's in any arrangement whatsoever. Here are all the possible arrangements:

SSFFF	SFSFF	SFFSF	SFFFS	FSSFF
FSFSF	FSFFS	FFSSF	FFSFS	FFFSS

There are 10 of them, all with the same probability. The overall probability of 2 successes is therefore

$$P(X = 2) = 10(0.25)^2(0.75)^3 = 0.26367$$

The pattern of this calculation works for any binomial probability. That is,

$$
\begin{aligned}
P(X = k) &= P(\text{exactly } k \text{ successes in } n \text{ trials}) \\
&= \text{number of arrangements} \cdot p^k(1 - p)^{n-k}
\end{aligned}
$$

To use this formula, we must count the number of arrangements of k successes in n observations. This number is called the **binomial coefficient**. We use the following fact to do the counting without actually listing all the arrangements.

DEFINITION: Binomial coefficient

The number of ways of arranging k successes among n observations is given by the **binomial coefficient**

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

for $k = 0, 1, 2, \ldots, n$ where

$$n! = n(n-1)(n-2) \cdot \ldots \cdot (3)(2)(1)$$

and $0! = 1$.

The formula for binomial coefficients uses the **factorial** notation. For any positive whole number n, its factorial $n!$ is

$$n! = n(n-1)(n-2) \cdot \ldots \cdot (3)(2)(1)$$

We also define $0! = 1$.

The larger of the two factorials in the denominator of a binomial coefficient will cancel much of the $n!$ in the numerator. For example, the binomial coefficient we need to find the probability that exactly 2 of the couple's 5 children inherit type O blood is

$$\binom{5}{2} = \frac{5!}{2!3!} = \frac{(5)(4)\cancel{(3)}\cancel{(2)}\cancel{(1)}}{(2)(1)\cancel{(3)}\cancel{(2)}\cancel{(1)}} = \frac{(5)(4)}{(2)(1)} = 10$$

Some people prefer the notation ${}_5C_2$ instead of $\binom{5}{2}$ for the binomial coefficient.

The binomial coefficient is $\binom{5}{2}$ *not related to the fraction* $\frac{5}{2}$. A helpful way to remember its meaning is to read it as "5 choose 2." Binomial coefficients have many uses, but we are interested in them only as an aid to finding binomial probabilities. If you need to compute a binomial coefficient, use your calculator.

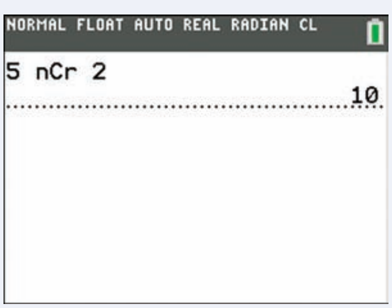
12. TECHNOLOGY CORNER BINOMIAL COEFFICIENTS ON THE CALCULATOR

TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

To calculate a binomial coefficient like $\binom{5}{2}$ on the TI-83/84 or TI-89, proceed as follows:

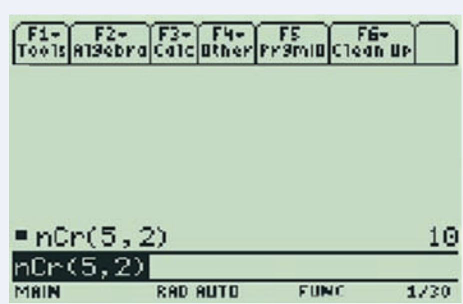
TI-83/84

- Type 5, press MATH, arrow over to PRB, choose `nCr`, and press ENTER. Then type 2 and press ENTER again to execute the command 5 `nCr` 2.



TI-89

- From the home screen, press 2nd 5 (MATH), choose `Probability, nCr(`, and press ENTER. Complete the command `nCr(5,2)` and press ENTER.



The binomial coefficient $\binom{n}{k}$ counts the number of different ways in which k successes can be arranged among n trials. The binomial probability $P(X = k)$ is this count multiplied by the probability of any one specific arrangement of the k successes. Here is the result we seek.

BINOMIAL PROBABILITY FORMULA

If X has the binomial distribution with n trials and probability p of success on each trial, the possible values of X are 0, 1, 2, . . . , n. If k is any one of these values,

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

With our formula in hand, we can now calculate any binomial probability.

EXAMPLE Inheriting Blood Type

Using the binomial probability formula

PROBLEM: Each child of a particular set of parents has probability 0.25 of having type O blood. Suppose the parents have 5 children.

(a) Find the probability that exactly 3 of the children have type O blood.

(b) Should the parents be surprised if more than 3 of their children have type O blood? Justify your answer.

SOLUTION: Let X = the number of children with type O blood. We know that X has a binomial distribution with $n = 5$ and $p = 0.25$.

(a) We want to find $P(X = 3)$.

$$P(X = 3) = \binom{5}{3}(0.25)^3(0.75)^2 = 10(0.25)^3(0.75)^2 = 0.08789$$

There is about a 9% chance that exactly 3 of the 5 children have type O blood.

(b) To answer this question, we need to find $P(X > 3)$.

$$P(X > 3) = P(X = 4) + P(X = 5) = \binom{5}{4}(0.25)^4(0.75)^1 + \binom{5}{5}(0.25)^5(0.75)^0$$
$$= 5(0.25)^4(0.75)^1 + 1(0.25)^5(0.75)^0 = 0.01465 + 0.00098 = 0.01563$$

Because there's only about a 1.5% chance of having more than 3 children with type O blood, the parents should definitely be surprised if this happens.

For Practice *Try Exercises* **75** *and* **77**

We could also use the calculator's `binompdf` and `binomcdf` commands to perform the calculations in the previous example. The following Technology Corner shows how to do it.

13. TECHNOLOGY CORNER BINOMIAL PROBABILITY ON THE CALCULATOR

TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

There are two handy commands on the TI-83/84 and TI-89 for finding binomial probabilities: `binompdf` and `binomcdf`. The inputs for both commands are the number of trials n, the success probability p, and the values of interest for the binomial random variable X.

`binompdf(n,p,k)` computes $P(X = k)$
`binomcdf(n,p,k)` computes $P(X \leq k)$

Let's use these commands to confirm our answers in the previous example.

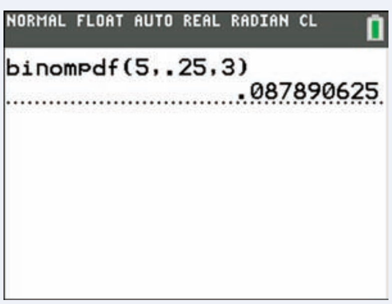
(a) Find the probability that exactly 3 of the children have type O blood.

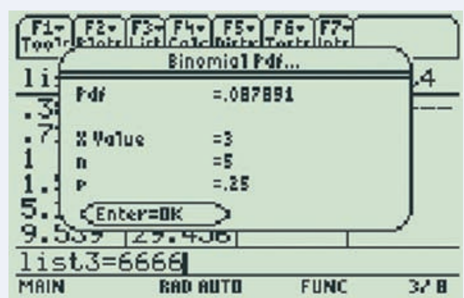
TI-83/84

- Press 2nd VARS (DISTR) and choose `binompdf(`.
- **OS 2.55 or later:** In the dialog box, enter these values: `trials:5, p:0.25, x value:3`, choose Paste, and then press ENTER. **Older OS:** Complete the command `binompdf(5,0.25,3)` and press ENTER.

TI-89

- In the Stats/List Editor, Press F5 (Distr) and choose `Binomial Pdf`.
- In the dialog box, enter these values: `Num Trials, n:5, Prob Success,p:0.25, X value:3`, and then choose ENTER.





These results agree with our previous answer using the binomial probability formula: 0.08789.

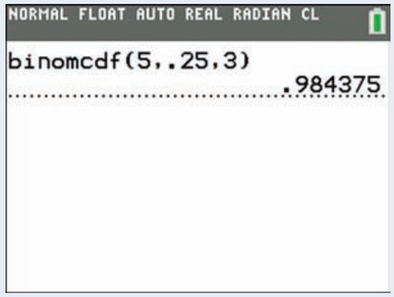
(b) Should the parents be surprised if more than 3 of their children have type O blood?
To find P(X > 3), use the complement rule:

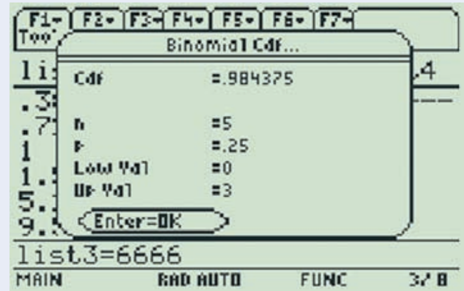
$$P(X > 3) = 1 - P(X \leq 3) = 1 - \texttt{binomcdf(5,0.25,3)}$$

- Press 2nd VARS (DISTR) and choose `binomcdf(`.
- **OS 2.55 or later:** In the dialog box, enter these values: `trials:5, p:0.25, x value:3`, choose Paste, and then press ENTER. Subtract this result from 1 to get the answer. **Older OS:** Complete the command `binomcdf(5,0.25,3)` and press ENTER. Subtract this result from 1 to get the answer.

- In the Stats/List Editor, Press F5 (Distr) and choose `Binomial Cdf`....
- In the dialog box, enter these values: `Num Trials, n:5, Prob Success,p:0.25, lower value:0, upper value:3` and then choose ENTER. Subtract this result from 1 to get the answer.

We could also have done the calculation for part (b) as $P(X > 3) = P(X = 4) + P(X = 5)$ = `binompdf(5,0.25,4)` + `binompdf(5,0.25,5)` = 0.01465 + 0.00098 = 0.01563.





Now we subtract from 1 to get the desired answer: $1 - 0.984375 = 0.015625$. This result agrees with our previous answer using the binomial probability formula: 0.01563.

AP® EXAM TIP Don't rely on "calculator speak" when showing your work on free-response questions. Writing `binompdf(5,0.25,3)` = 0.08789 will *not* earn you full credit for a binomial probability calculation. At the very least, you must indicate what each of those calculator inputs represents. For example, "I used binompdf(trials:5,p:0.25,x value:3)."

Note the use of the complement rule to find $P(X > 3)$ in the Technology Corner: $P(X > 3) = 1 - P(X \leq 3)$. This is necessary because the calculator's `binomcdf(n,p,k)` command only computes the probability of getting k *or fewer* successes in n trials. Students often have trouble identifying the correct third input for the `binomcdf` command when a question asks them to find the probability of getting less than, more than, or at least so many successes.

Here's a helpful tip to avoid making such a mistake: write out the possible values of the variable, circle the ones you want to find the probability of, and cross out the rest. In the previous example, X can take values from 0 to 5 and we want to find $P(X > 3)$:

$$\cancel{0} \quad \cancel{1} \quad \cancel{2} \quad \cancel{3} \quad 4 \quad 5$$

Crossing out the values from 0 to 3 shows why the correct calculation is $1 - P(X \leq 3)$.

Take another look at the solution in the blood-type example. The structure is much like the one we used when doing Normal calculations. Here is a revised summary box that describes the process.

HOW TO FIND BINOMIAL PROBABILITIES

Step 1: State the distribution and the values of interest. Specify a binomial distribution with the number of trials n, success probability p, and the values of the variable clearly identified.

Step 2: Perform calculations—show your work! Do one of the following: (i) Use the binomial probability formula to find the desired probability; or (ii) use the `binompdf` or `binomcdf` command and label each of the inputs.

Step 3: Answer the question.

Here's an example that shows the method at work.

EXAMPLE Free Lunch?

Binomial calculations

A local fast-food restaurant is running a "Draw a three, get it free" lunch promotion. After each customer orders, a touch-screen display shows the message "Press here to win a free lunch." A computer program then simulates one card being drawn from a standard deck. If the chosen card is a 3, the customer's order is free. Otherwise, the customer must pay the bill.

PROBLEM:

(a) All 12 players on a school's basketball team place individual orders at the restaurant. What is the probability that exactly 2 of them win a free lunch?

(b) If 250 customers place lunch orders on the first day of the promotion, what's the probability that fewer than 10 win a free lunch?

SOLUTION:

(a) **Step 1: State the distribution and the values of interest.** Let X = the number of players who win a free lunch. There are 12 independent trials of the chance process, each with success probability 4/52 (because there are 4 threes in a standard deck of 52 cards). So X has a binomial distribution with $n = 12$ and $p = 4/52$. We want to find $P(X = 2)$.

Step 2: Perform calculations—show your work! The binomial probability formula gives

$$P(X = 2) = \binom{12}{2}\left(\frac{4}{52}\right)^2\left(\frac{48}{52}\right)^{10} = 0.1754$$

Using technology: The command `binompdf(trials:12,p:4/52,x value:2)` also gives 0.1754.

Step 3: Answer the question. There is about a 17.5% probability that exactly 2 players will win a free lunch.

(b) **Step 1: State the distribution and the values of interest.** Let Y = the number of customers who win a free lunch. There are 250 independent trials of the chance process, each with success probability 4/52. So Y has a binomial distribution with $n = 250$ and $p = 4/52$. We want to find $P(Y < 10)$.

Step 2: Perform calculations—show your work! The values of Y that interest us are

$$0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 10 \quad 11 \quad 12 \quad \ldots \quad 250$$

To use the binomial formula, we would have to add up the probabilities for $Y = 0, Y = 1, \ldots, Y = 9$. That's too much work! The better option is to use technology:

$P(Y < 10) = P(Y \leq 9) =$ `binomcdf(trials:250,p:4/52,x value:9)`
$= 0.00613$

Step 3: Answer the question. There is almost no chance that fewer than 10 customers will win a free lunch. If this actually happened, the customers should be suspicious about the restaurant's claim.

For Practice *Try Exercise* 79

CHECK YOUR UNDERSTANDING

To introduce her class to binomial distributions, Mrs. Desai gives a 10-item, multiple-choice quiz. The catch is, students must simply guess an answer (A through E) for each question. Mrs. Desai uses her computer's random number generator to produce the answer key, so that each possible answer has an equal chance to be chosen. Patti is one of the students in this class. Let X = the number of Patti's correct guesses.

1. Show that X is a binomial random variable.
2. Find $P(X = 3)$. Explain what this result means.
3. To get a passing score on the quiz, a student must guess correctly at least 6 times. Would you be surprised if Patti earned a passing score? Compute an appropriate probability to support your answer.

Mean and Standard Deviation of a Binomial Distribution

What does the probability distribution of a binomial random variable look like? The table below shows the possible values and corresponding probabilities for X = the number of children with type O blood. This is a binomial random variable with $n = 5$ and $p = 0.25$. Figure 6.14 shows a histogram of the probability distribution.

Value x_i:	0	1	2	3	4	5
Probability p_i:	0.23730	0.39551	0.26367	0.08789	0.01465	0.00098

Let's describe what we see.

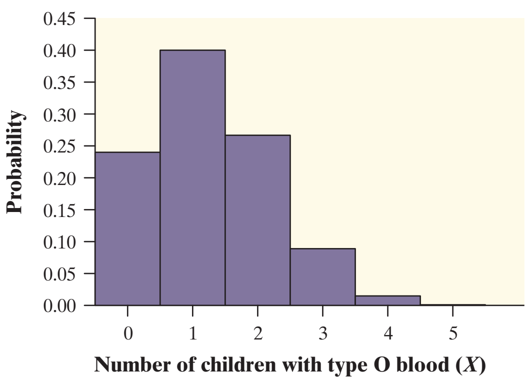


FIGURE 6.14 Histogram showing the probability distribution of the binomial random variable X = number of children with type O blood in a family with 5 children.

Shape: The probability distribution of X is skewed to the right. Because the chance that any one of the couple's children inherits type O blood is 0.25, it's quite likely that 0, 1, or 2 of the children will have type O blood. Larger values of X are much less likely.

Center: The median number of children with type O blood is 1 because that's where the 50th percentile of the distribution falls. How about the mean? It's

$$\mu_X = \sum x_i p_i = (0)(0.23730) + (1)(0.39551) + \cdots + (5)(0.00098) = 1.25$$

So the expected number of children with type O blood in families like this one with 5 children is 1.25.

Spread: The variance of X is

$$\begin{aligned}\sigma_X^2 &= \sum (x_i - \mu_X)^2 p_i \\ &= (0 - 1.25)^2(0.23730) + (1 - 1.25)^2(0.39551) + \cdots + \\ &\quad (5 - 1.25)^2(0.00098) = 0.9375\end{aligned}$$

So the standard deviation of X is

$$\sigma_X = \sqrt{0.9375} = 0.968$$

The number of children with type O blood will typically differ from the mean by about 0.968 in families like this one with 5 children.

Did you think about why the mean is $\mu_X = 1.25$? Because each child has a 0.25 chance of inheriting type O blood, we'd expect one-fourth of the 5 children to have this blood type. In other words, $\mu_X = 5(0.25) = 1.25$. This method can be used to find the mean of any binomial random variable with parameters n and p:

$$\mu_X = np$$

There are fairly simple formulas for the variance and standard deviation, too, but they aren't as easy to explain:

$$\sigma_X^2 = np(1-p) \quad \text{and} \quad \sigma_X = \sqrt{np(1-p)}$$

For our family with 5 children,

$$\sigma_X^2 = 5(0.25)(0.75) = 0.9375 \quad \text{and} \quad \sigma_X = \sqrt{0.9375} = 0.968$$

MEAN AND STANDARD DEVIATION OF A BINOMIAL RANDOM VARIABLE

If a count X of successes has the binomial distribution with number of trials n and probability of success p, the **mean** and **standard deviation** of X are

$$\mu_X = np$$
$$\sigma_X = \sqrt{np(1-p)}$$

Remember that these formulas work only for binomial distributions. They can't be used for other distributions. Caution

THINK ABOUT IT

Where do the binomial mean and variance formulas come from? We can derive the formulas for the mean and variance of a binomial distribution using what we learned about combining random variables in Section 6.2. Let's start with the random variable B that's described by the following probability distribution:

Value b_i:	0	1
Probability p_i:	$1-p$	p

You can think of B as representing the result of a single trial of some chance process. If a success occurs (probability p), then $B = 1$. If a failure occurs, then $B = 0$. Notice that the mean of B is

$$\mu_B = \sum b_i p_i = (0)(1-p) + (1)(p) = p$$

and that the variance of B is

$$\begin{aligned}\sigma_B^2 = \sum(b_i - \mu_B)^2 p_i &= (0-p)^2(1-p) + (1-p)^2 p \\ &= p(1-p)[p + (1-p)] \\ &= p(1-p)\end{aligned}$$

Now consider the random variable $X = B_1 + B_2 + \cdots + B_n$. We can think of X as counting the number of successes in n independent trials of this chance process, with each trial having success probability p. In other words, X is a bino-

mial random variable with parameters n and p. By the rules from Section 6.2, the mean of X is

$$\mu_X = \mu_{B_1} + \mu_{B_2} + \cdots + \mu_{B_n} = p + p + \cdots + p = np$$

and the variance of X is

$$\begin{aligned}\sigma_X^2 &= \sigma_{B_1}^2 + \sigma_{B_2}^2 + \cdots + \sigma_{B_n}^2\\ &= p(1-p) + p(1-p) + \cdots + p(1-p)\\ &= np(1-p)\end{aligned}$$

The standard deviation of X is therefore

$$\sigma_X = \sqrt{np(1-p)}$$

EXAMPLE Bottled Water versus Tap Water

Binomial distribution in action

Mr. Bullard's AP® Statistics class did the Activity on page 346. There were 21 students in the class. If we assume that the students in his class could *not* tell tap water from bottled water, then each one was basically guessing, with a 1/3 chance of being correct. Let X = the number of students who correctly identified the cup containing the different type of water.

PROBLEM:

(a) Explain why X is a binomial random variable.

(b) Find the mean and standard deviation of X. Interpret each value in context.

(c) Of the 21 students in the class, 13 made correct identifications. Are you convinced that Mr. Bullard's students could tell bottled water from tap water? Justify your answer.

SOLUTION:

(a) Assuming that students were just guessing, the Activity consisted of 21 repetitions of this chance process. Let's check the BINS:

- Binary? On each trial, "success" = correct identification; "failure" = incorrect identification.
- Independent? One student's result should tell us nothing about any other student's result.
- Number? There were $n = 21$ trials.
- Success? Each student had a $p = 1/3$ chance of guessing correctly.

Because X is counting the number of successes in this binomial setting, it is a binomial random variable.

(b) The mean of X is

$$\mu_X = np = 21(1/3) = 7$$

If the Activity were repeated many times with groups of 21 students who were just guessing, the average number of students who guess correctly would be about 7. The standard deviation of X is

$$\sigma_X = \sqrt{np(1-p)} = \sqrt{21(1/3)(2/3)} = 2.16$$

If the Activity were repeated many times with groups of 21 students who were just guessing, the number of correct identifications would typically differ from 7 by about 2.16.

(c) The class's result corresponds to $X = 13$, a value that's nearly 3 standard deviations above the mean. How likely is it that 13 or more of Mr. Bullard's students would guess correctly? It's

$$P(X \geq 13) = 1 - P(X \leq 12)$$

Using the calculator's `binomcdf(trials:21,p:1/3,xvalue:12)` command:

$$P(X \geq 13) = 1 - 0.9932 = 0.0068$$

The students had less than a 1% chance of getting so many right if they were all just guessing. This is strong evidence that some of the students in the class could tell bottled water from tap water.

For Practice *Try Exercise* 85

Figure 6.15 shows the probability distribution for the number of correct guesses in Mr. Bullard's class if no one can tell bottled water from tap water. As you can see from the graph, the chance of 13 or more guessing correctly is quite small.

Although the histogram is slightly right-skewed (there's a long tail that extends out to $X = 21$), it looks like a Normal density curve might fit the bulk of the distribution fairly well.

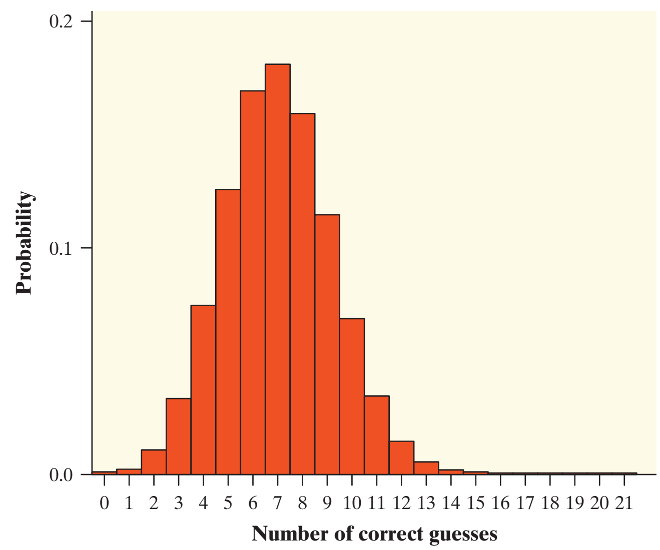


FIGURE 6.15 Histogram of the probability distribution for the binomial random variable X = number of correct guesses in Mr. Bullard's class.

CHECK YOUR UNDERSTANDING

Refer to the previous Check Your Understanding (page 397) about Mrs. Desai's special multiple-choice quiz on binomial distributions. We defined X = the number of Patti's correct guesses.

1. Find μ_X. Interpret this value in context.
2. Find σ_X. Interpret this value in context.
3. What's the probability that the number of Patti's correct guesses is more than 2 standard deviations above the mean? Show your method.

Binomial Distributions in Statistical Sampling

The binomial distributions are important in statistics when we wish to make inferences about the proportion p of successes in a population. Here is an example involving a familiar product.

EXAMPLE Bad Flash Drives

Binomial distributions and sampling

A supplier inspects an SRS of 10 flash drives from a shipment of 10,000 flash drives. Suppose that (unknown to the supplier) 2% of the flash drives in the shipment are defective. Count the number X of bad flash drives in the sample.

This is not quite a binomial setting. Removing 1 flash drive changes the proportion of bad flash drives remaining in the shipment. The conditional probability that the second flash drive chosen is bad changes when we know whether the first is good or bad. But removing 1 flash drive from a shipment of 10,000 changes the makeup of the remaining 9999 flash drives very little. The distribution of X is very close to the binomial distribution with $n = 10$ and $p = 0.02$. To illustrate this, let's compute the probability that none of the 10 flash drives is defective. Using the binomial distribution, it's

$$P(X = 0) = \binom{10}{0}(0.02)^0(0.98)^{10} = 0.8171$$

The actual probability of getting no defective flash drives is

$$P(\text{no defectives}) = \frac{9800}{10{,}000} \times \frac{9799}{9999} \times \frac{9798}{9998} \times \cdots \times \frac{9791}{9991} = 0.8170$$

Those two probabilities are quite close!

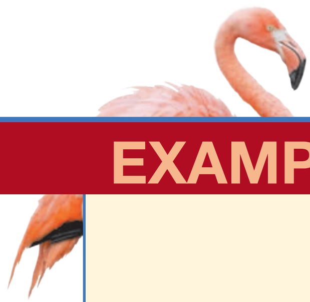
Almost all real-world sampling, such as taking an SRS from a population of interest, is done without replacement. As the previous example illustrates, sampling without replacement leads to a violation of the independence condition.

The flash drives example shows how we can use binomial distributions in the statistical setting of selecting an SRS. When the population is much larger than the sample, a count of successes in an SRS of size n has approximately the binomial distribution with n equal to the sample size and p equal to the proportion of successes in the population. What counts as "much larger"? In practice, the binomial distribution gives a good approximation as long as we don't sample more than 10% of the population. We refer to this as the **10% condition**.

10% CONDITION

When taking an SRS of size n from a population of size N, we can use a binomial distribution to model the count of successes in the sample as long as $n \leq \frac{1}{10}N$.

Here's an example that shows why it's important to check the 10% condition before calculating a binomial probability. You might recognize the setting from the first activity in the book (page 5).



EXAMPLE Hiring Discrimination—It Just Won't Fly!

Sampling without replacement

An airline has just finished training 25 first officers—15 male and 10 female—to become captains. Unfortunately, only eight captain positions are available right now. Airline managers decide to use a lottery to determine which pilots will fill the available positions. Of the 8 captains chosen, 5 are female and 3 are male.

PROBLEM: Explain why the probability that 5 female pilots are chosen in a fair lottery is *not*

$$P(X=5)=\binom{8}{5}(0.40)^5(0.60)^3=0.124$$

(The correct probability is 0.106.)

SOLUTION: The managers are sampling without replacement when they do the lottery. There's a 0.40 chance that the first pilot selected for a captain position is female. Once that person is chosen, the probability that the next pilot selected will be female is no longer 0.40. The binomial formula assumes that the conditional probability of success stays constant at 0.40 throughout the eight trials of this chance process. This calculation will be approximately correct if the success probability doesn't change too much—as long as we don't sample more than 10% of the population. In this case, managers are sampling 8 out of 25 pilots—almost 1/3 of the population. That explains why the binomial probability is off by about 17% (0.018/0.106) from the correct answer.

For Practice *Try Exercise* 87

The Normal approximation to binomial distributions* As n gets larger, something interesting happens to the shape of the binomial distribution. Figure 6.16 shows histograms of binomial distributions for different values of n and p. As the number of observations n becomes larger, the binomial distribution gets close to a Normal distribution. You can investigate the relationship between n and p yourself using the *Normal Approximation to Binomial Distributions* applet at the book's Web site.

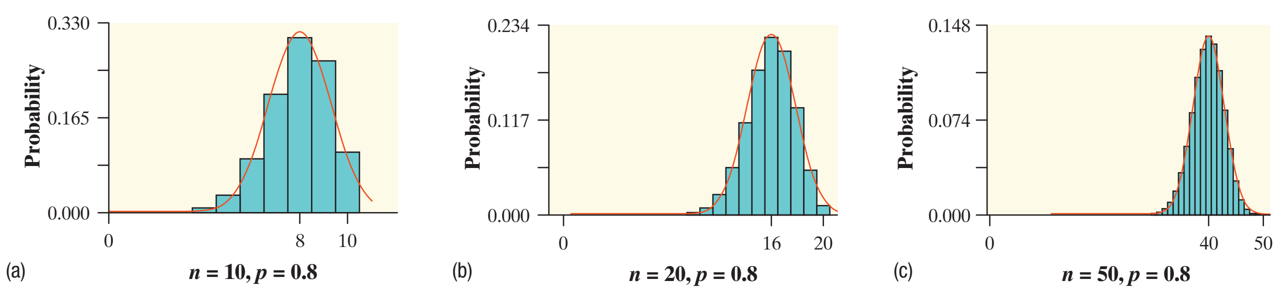


FIGURE 6.16 Histograms of binomial distributions with (a) $n = 10$ and $p = 0.8$, (b) $n = 20$ and $p = 0.8$, and (c) $n = 50$ and $p = 0.8$. As n increases, the shape of the probability distribution gets more and more Normal.

*This topic is not required for the AP® Statistics exam. Some teachers prefer to discuss this topic when presenting the sampling distribution of $\hat{p}$ (Chapter 7).

When n is large, we can use Normal probability calculations to approximate binomial probabilities. Here are the facts.

NORMAL APPROXIMATION FOR BINOMIAL DISTRIBUTIONS: THE LARGE COUNTS CONDITION

Suppose that a count X of successes has the binomial distribution with n trials and success probability p. When n is large, the distribution of X is approximately Normal with

$$\text{mean: } \mu_X = np \quad \text{and} \quad \text{standard deviation: } \sigma_X = \sqrt{np(1-p)}$$

As a rule of thumb, we will use the Normal approximation when n is so large that

$$np \geq 10 \qquad \text{and} \qquad n(1-p) \geq 10$$

That is, the expected number of successes and failures are both at least 10. We refer to this as the **Large Counts condition**.

The Normal approximation is easy to remember because it says to act as if X is Normal with exactly the same mean and standard deviation as the binomial. The accuracy of the Normal approximation improves as the sample size n increases. It is most accurate for any fixed n when p is close to 1/2 and least accurate when p is near 0 or 1. This is why the rule of thumb in the box depends on p as well as n.

EXAMPLE Attitudes toward Shopping

Normal approximation to a binomial

Sample surveys show that fewer people enjoy shopping than in the past. A survey asked a nationwide random sample of 2500 adults if they agreed or disagreed that "I like buying new clothes, but shopping is often frustrating and time-consuming."[11] The population that the poll wants to draw conclusions about is all U.S. residents aged 18 and over.

PROBLEM: Suppose that exactly 60% of all adult U.S. residents would say "Agree" if asked the same question. Let X = the number in the sample who agree.

(a) Show that X is approximately a binomial random variable.

(b) Check the conditions for using a Normal approximation in this setting.

(c) Use a Normal distribution to estimate the probability that 1520 or more of the sample agree.

SOLUTION:

(a) Let's check the BINS.

- Binary? Success = agree that shopping is frustrating, failure = don't agree.
- Independent? The trials are not independent: the conditional probability of a success changes due to the sampling without replacement. But the *10% condition* is met because 2500 people is much less than 10% of all U.S. adult residents.
- Number? There are $n = 2500$ trials of this chance process.
- Success? There is the same probability of selecting an adult who agrees on each trial: $p = 0.6$.

So the number in our sample who agree that shopping is frustrating is a random variable X having roughly the binomial distribution with $n = 2500$ and $p = 0.6$.

(b) We need to check the Large Counts condition. Because $np = 2500(0.6) = 1500$ and $n(1 - p) = 2500(0.4) = 1000$ are both at least 10, we should be safe using the Normal approximation.

(c) **Step 1: State the distribution and the values of interest.** Act as though the count X has the Normal distribution with the same mean and standard deviation as the binomial distribution:

$$\mu = np = 2500(0.6) = 1500 \quad \text{and} \quad \sigma = \sqrt{np(1-p)} = \sqrt{2500(0.6)(0.4)} = 24.49$$

We want to find $P(X \geq 1520)$. Figure 6.17 shows this probability as the area under a Normal curve.

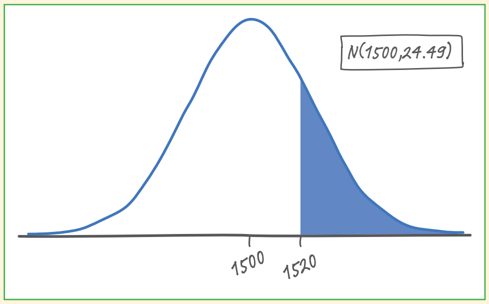


FIGURE 6.17 Normal distribution to approximate the binomial probability of getting 1520 or more successes when $n = 2500$ and $p = 0.6$.

Step 2: Perform calculations—show your work! Standardizing the boundary value gives

$$z = \frac{1520 - 1500}{24.49} = 0.82$$

Using Table A, the probability we want is

$$P(Z \geq 0.82) = 1 - 0.7939 = 0.2061$$

Using technology: The command `normalcdf(lower:1520, upper:10000, μ:1500, σ:24.49)` gives an area of 0.2071.

Step 3: Answer the question. There is about a 21% chance of getting a sample in which 1520 or more agree with the statement.

For Practice *Try Exercise* **91**

We can also find the probability that 1520 or more of the sample agree that shopping is often frustrating and time-consuming using the command `1-binomcdf(2500,0.6,1519)`, which yields 0.2131. The Normal approximation, 0.2061, misses the more accurate binomial probability by about 0.007.

Geometric Random Variables

In a binomial setting, the number of trials n is fixed in advance, and the binomial random variable X counts the number of successes. The possible values of X are 0, 1, 2, . . . , n. In other situations, the goal is to repeat a chance process *until a success occurs:*

- Roll a pair of dice until you get doubles.
- In basketball, attempt a three-point shot until you make one.
- Keep placing a $1 bet on the number 15 in roulette until you win.

These are all examples of a **geometric setting.** Although the number of trials isn't fixed in advance, the trials are independent and the probability of success remains constant.

DEFINITION: Geometric setting

A **geometric setting** arises when we perform independent trials of the same chance process and record the number of trials it takes to get one success. On each trial, the probability p of success must be the same.

Here's an Activity your class can try that involves a geometric setting.

ACTIVITY | Is This Your Lucky Day?

MATERIALS:
Calculator or computer random number generator to select student names and days of the week

Your teacher is planning to give you 10 problems for homework. As an alternative, you can agree to play the Lucky Day Game. Here's how it works. A student will be selected at random from your class and asked to pick a day of the week (for instance, Thursday). Then your teacher will use technology to randomly choose a day of the week as the "lucky day." If the student picks the correct day, the class will have only one homework problem. If the student picks the wrong day, your teacher will select another student from the class at random. The chosen student will pick a day of the week and your teacher will use technology to choose a "lucky day." If this student gets it right, the class will have two homework problems. The game continues until a student correctly guesses the lucky day. Your teacher will assign a number of homework problems that is equal to the total number of guesses made by members of your class. Are you ready to play the Lucky Day Game?

1. Decide as a class about whether to "gamble" on the number of homework problems you will receive. You have 30 seconds.
2. Play the Lucky Day Game and see what happens!

In a geometric setting, if we define the random variable Y to be the number of trials needed to get the first success, then Y is called a **geometric random variable.** The probability distribution of Y is a **geometric distribution.**

DEFINITION: Geometric random variable and geometric distribution

The number of trials Y that it takes to get a success in a geometric setting is a **geometric random variable**. The probability distribution of Y is a **geometric distribution** with parameter p, the probability of a success on any trial. The possible values of Y are 1, 2, 3, ...

As with binomial random variables, it's important to be able to distinguish situations in which a geometric distribution does and doesn't apply.

EXAMPLE The Lucky Day Game

Geometric settings and random variables

The random variable of interest in this game is Y = the number of picks it takes to correctly match the lucky day. Each pick is one trial of the chance process. Knowing the result of one student's pick tells us nothing about the result of any other pick. On each trial, the probability of a correct pick is 1/7.

This is a geometric setting. Because Y counts the number of trials to get the first success, it is a geometric random variable with parameter $p = 1/7$.

What is the probability that the first student picks correctly and wins the Lucky Day Game? It's $P(Y = 1) = 1/7$. That's also the class's chance of getting only one homework problem. For the class to have two homework problems, the first student selected must pick an incorrect day of the week and the second student must pick the lucky day correctly. The probability that this happens is $P(Y = 2) = (6/7)(1/7) = 0.1224$. Likewise, $P(Y = 3) = (6/7)(6/7)(1/7) = 0.1050$. In general, the probability that the first correct pick comes on the kth trial is $P(Y = k) = (6/7)^{k-1}(1/7)$. Let's summarize what we've learned about calculating a **geometric probability.**

GEOMETRIC PROBABILITY FORMULA

If Y has the geometric distribution with probability p of success on each trial, the possible values of Y are 1, 2, 3, ... If k is any one of these values,

$$P(Y = k) = (1 - p)^{k-1}p$$

With our formula in hand, we can now compute any geometric probability.

EXAMPLE The Lucky Day Game

Calculating geometric probabilities

PROBLEM: Let the random variable Y be defined as in the previous example.

(a) Find the probability that the class receives exactly 10 homework problems as a result of playing the Lucky Day Game.

(b) Find $P(Y < 10)$ and interpret this value in context.

SOLUTION: Y = the number of attempts it takes to get a correct pick = the number of homework problems.

(a) $P(Y = 10) = (6/7)^9(1/7) = 0.0357$.

(b) $P(Y < 10) = P(Y = 1) + P(Y = 2) + P(Y = 3) + \ldots + P(Y = 9) = 1/7 + (6/7)(1/7) + (6/7)^2(1/7) + \ldots + (6/7)^8(1/7) = 0.7503$. There's about a 75% chance that the class will get less homework by playing the Lucky Day Game.

For Practice *Try Exercise* 97

As you probably guessed, we used the calculator's `geometpdf` and `geometcdf` commands for the computations in the previous example. The following Technology Corner shows you how we did it.

14. TECHNOLOGY CORNER GEOMETRIC PROBABILITY ON THE CALCULATOR

TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

There are two handy commands on the TI-83/84 and TI-89 for finding geometric probabilities: `geometpdf` and `geometcdf`. The inputs for both commands are the success probability p and the value(s) of interest for the geometric random variable Y.

`geometpdf(p,k)` computes $P(Y = k)$

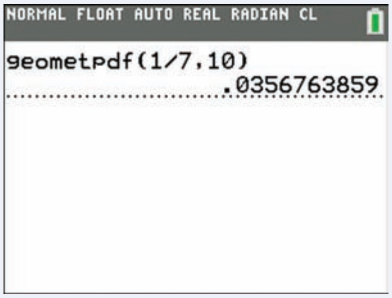
`geometcdf(p,k)` computes $P(Y \leq k)$

Let's use these commands to confirm our answers in the previous example.

(a) Find the probability that the class receives exactly 10 homework problems as a result of playing the Lucky Day Game.

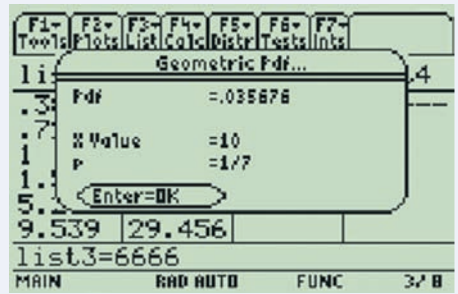
TI-83/84

- Press 2nd VARS (DISTR) and choose `geometpdf(`.
- **OS 2.55 or later:** In the dialog box, enter these values: `p:1/7`, `x value:10`, choose Paste, and then press ENTER. **Older OS:** Complete the command `geometpdf(1/7,10)` and press ENTER.



TI-89

- In the Stats/List Editor, Press F5 (Distr) and choose `Geometric Pdf...`.
- In the dialog box, enter these values: `Prob Success, p:1/7, X value:10`, and then choose ENTER.



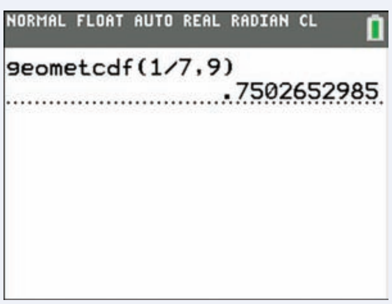
These results agree with our previous answer using the geometric probability formula: 0.0357.

(b) Find $P(Y < 10)$ and interpret this value in context. To find $P(Y < 10)$, use the `geometcdf` command:

$$P(Y < 10) = P(Y \leq 9) = \texttt{geometcdf(1/7,9)}$$

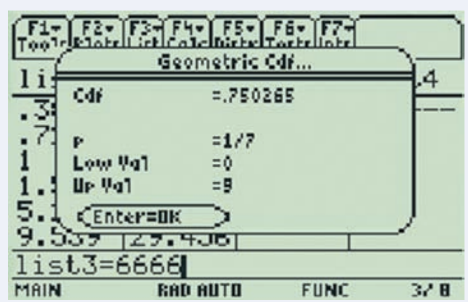
TI-83/84

- Press 2nd VARS (DISTR) and choose `geometcdf(`.
- **OS 2.55 or later:** In the dialog box, enter these values: `p:1/7`, `x value:9`, choose `Paste`, and then press ENTER. **Older OS:** Complete the command `geometcdf(1/7,9)` and press ENTER.



TI-89

- In the Stats/List Editor, Press F5 (Distr) and choose `Geometric Cdf....`
- In the dialog box, enter these values: `Prob Success, p:1/7, Lower value:0, Upper value:9`, and then choose ENTER.



These results agree with our previous answer using the geometric probability formula: 0.7503.

The table below shows part of the probability distribution of Y. We can't show the entire distribution, because the number of trials it takes to get the first success could be a very large number.

Value y_i:	1	2	3	4	5	6	7	8	9	. . .
Probability p_i:	0.143	0.122	0.105	0.090	0.077	0.066	0.057	0.049	0.042	

Figure 6.18 is a histogram of the probability distribution for values of Y from 1 to 26. Let's describe what we see.

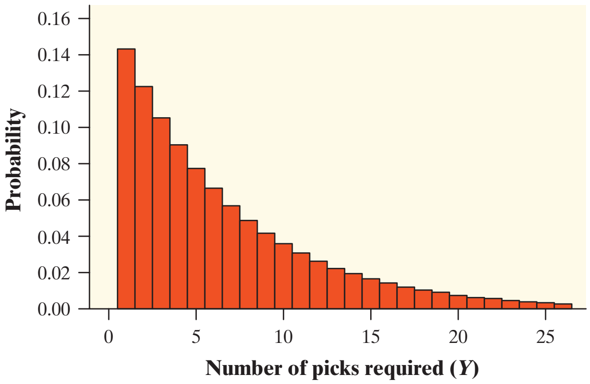


FIGURE 6.18 Histogram showing the probability distribution of the geometric random variable Y = number of trials needed for students to pick correctly in the Lucky Day Game.

Shape: The heavily right-skewed shape is characteristic of any geometric distribution. That's because the most likely value of a geometric random variable is 1. The probability of each successive value decreases by a factor of $(1 - p)$.

Center: The mean of Y is $\mu_Y = 7$. (Due to the infinite number of possible values of Y, the calculation of the mean is beyond the scope of this text.) If the class played the Lucky Day Game many times, they would receive an average of 7 homework problems. It's no coincidence that $p = 1/7$ and $\mu_Y = 7$. With probability of success 1/7 on each trial, we'd expect it to take an average of 7 trials to get the first success.

Spread: The standard deviation of Y is $\sigma_Y = 6.48$. (Due to the infinite number of possible values of Y, the calculation of the standard deviation is beyond the scope of this text.) If the class played the Lucky Day game many times, the number of homework problems they receive would typically differ from 7 by about 6.5 problems. That could mean a lot of homework!

We can generalize the result for the mean of a geometric random variable.

MEAN (EXPECTED VALUE) OF A GEOMETRIC RANDOM VARIABLE

If Y is a geometric random variable with probability of success p on each trial, then its **mean** (expected value) is $\mu_Y = E(Y) = \frac{1}{p}$. That is, the expected number of trials required to get the first success is $1/p$.

CHECK YOUR UNDERSTANDING

Suppose you roll a pair of fair, six-sided dice until you get doubles. Let T = the number of rolls it takes.

1. Show that T is a geometric random variable.
2. Find $P(T = 3)$. Interpret this result in context.
3. In the game of Monopoly, a player can get out of jail free by rolling doubles within 3 turns. Find the probability that this happens.

case closed

A Jury of Your Peers?

In the chapter-opening Case Study on page 345, a defense attorney challenged the jury-pool selection process in his accused client's trial. Here are the facts:

- About 7.28% of the citizens in the court's jurisdiction were black.
- The jury pool had between 60 and 100 members, 3 of whom were black.

Use what you have learned in this chapter to help answer the following questions.

For now, assume that the court carried out a proper random-selection process to obtain a jury pool with 100 members.

1. Let X = the number of black citizens in the jury pool. What distribution does the random variable X have? Justify your answer.
2. Find the mean and standard deviation of X. Interpret these values in context.
3. If a jury pool has 3 or fewer blacks, should we be suspicious that the court did not carry out the random selection process correctly? Compute $P(X \leq 3)$ and use this result to support your answer.

What if the jury pool had 60 members? Assume once again that the court carried out a proper random-selection process. Let Y = the number of black citizens in the jury pool.

4. Without doing any calculations, decide if $P(Y \leq 3)$ is greater than, equal to, or less than $P(X \leq 3)$. Justify your answer.
5. Using the logic of Question 4, explain why you do not have to consider jury pools with 61, 62, . . . , 99 members to render a verdict about whether or not the jury-selection process was carried out properly. What is your verdict?

Section 6.3 Summary

- A **binomial setting** consists of n independent trials of the same chance process, each resulting in a success or a failure, with probability of success p on each trial. Remember to check the BINS! The count X of successes is a **binomial random variable.** Its probability distribution is a **binomial distribution.**
- The **binomial coefficient**

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

counts the number of ways k successes can be arranged among n trials. The **factorial** of n is

$$n! = n(n-1)(n-2) \cdot \ldots \cdot (3)(2)(1)$$

for positive whole numbers n, and $0! = 1$.
- If X has the binomial distribution with parameters n and p, the possible values of X are the whole numbers 0, 1, 2, . . . , n. The **binomial probability** of observing k successes in n trials is

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Binomial probabilities are best found using technology.

- The **mean** and **standard deviation** of a binomial random variable X are

$$\mu_X = np \qquad \sigma_X = \sqrt{np(1-p)}$$

- The binomial distribution with n trials and probability p of success gives a good approximation to the count of successes in an SRS of size n from a large population containing proportion p of successes. This is true as long as the sample size n is no more than 10% of the population size N (the **10% condition**).
- The **Normal approximation** to the binomial distribution* says that if X is a count of successes having the binomial distribution with parameters n and p, then when n is large, X is approximately Normally distributed with mean np and standard deviation $\sqrt{np(1-p)}$. We will use this approximation when $np \geq 10$ and $n(1-p) \geq 10$ (the **Large Counts condition**).
- A **geometric setting** consists of repeated trials of the same chance process in which the probability p of success is the same on each trial, and the goal is to count the number of trials it takes to get one success. If Y = the number of trials required to obtain the first success, then Y is a **geometric random variable.** Its probability distribution is called a **geometric distribution.**
- If Y has the geometric distribution with probability of success p, the possible values of Y are the positive integers 1, 2, 3, ... The **geometric probability** that Y takes any value is

$$P(Y = k) = (1-p)^{k-1}p$$

- The **mean** (expected value) of a geometric random variable Y is $\mu_Y = 1/p$.

*This topic is not required for the AP® Statistics exam. Some teachers prefer to discuss this topic when presenting the sampling distribution of $\hat{p}$ (Chapter 7).

6.3 TECHNOLOGY CORNERS

TI-Nspire Instructions in Appendix B; HP Prime instructions on the book's Web site.

12. Binomial coefficients on the calculator — page 392
13. Binomial probability on the calculator — page 394
14. Geometric probability on the calculator — page 406

Section 6.3 Exercises

In Exercises 69 to 72, determine whether the given random variable has a binomial distribution. Justify your answer.

69. **Sowing seeds** Seed Depot advertises that its new flower seeds have an 85% chance of germinating (growing). Suppose that the company's claim is true. Judy gets a packet with 20 randomly selected new flower seeds from Seed Depot and plants them in her garden. Let X = the number of seeds that germinate.

pg 388

70. **Long or short?** Put the names of all the students in your class in a hat. Mix them up, and draw four names without looking. Let Y = the number whose last names have more than six letters.

71. **Lefties** Exactly 10% of the students in a school are left-handed. Select students at random from the school, one at a time, until you find one who is left-handed. Let V = the number of students chosen.

pg 388

72. **Taking the train** According to New Jersey Transit, the 8:00 A.M. weekday train from Princeton to New York City has a 90% chance of arriving on time on a randomly selected day. Suppose this claim is true. Choose 6 days at random. Let W = the number of days on which the train arrives late.

73. **Binomial setting?** A binomial distribution will be approximately correct as a model for one of these two settings and not for the other. Explain why by briefly discussing both settings.

(a) When an opinion poll calls residential telephone numbers at random, only 20% of the calls reach a person. You watch the random digit dialing machine make 15 calls. X is the number that reach a person.

(b) When an opinion poll calls residential telephone numbers at random, only 20% of the calls reach a live person. You watch the random digit dialing machine make calls. Y is the number of calls needed to reach a live person.

74. **Binomial setting?** A binomial distribution will be approximately correct as a model for one of these two sports settings and not for the other. Explain why by briefly discussing both settings.

(a) A National Football League kicker has made 80% of his field goal attempts in the past. This season he attempts 20 field goals. The attempts differ widely in distance, angle, wind, and so on.

(b) A National Basketball Association player has made 80% of his free-throw attempts in the past. This season he takes 150 free throws. Basketball free throws are always attempted from 15 feet away with no interference from other players.

75. **Elk** Biologists estimate that a baby elk has a 44% chance of surviving to adulthood. Assume this estimate is correct. Suppose researchers choose 7 baby elk at random to monitor. Let X = the number who survive to adulthood. Use the binomial probability formula to find $P(X = 4)$. Interpret this result in context.

pg 393

76. **Rhubarb** Suppose you purchase a bundle of 10 bare-root rhubarb plants. The sales clerk tells you that 5% of these plants will die before producing any rhubarb. Assume that the bundle is a random sample of plants and that the sales clerk's statement is accurate. Let Y = the number of plants that die before producing any rhubarb. Use the binomial probability formula to find $P(Y = 1)$. Interpret this result in context.

77. **Elk** Refer to Exercise 75. How surprising would it be for more than 4 elk in the sample to survive to adulthood? Calculate an appropriate probability to support your answer.

pg 393

78. **Rhubarb** Refer to Exercise 76. Would you be surprised if 3 or more of the plants in the bundle die before producing any rhubarb? Calculate an appropriate probability to support your answer.

79. **Sowing seeds** Refer to Exercise 69.

pg 396

(a) Find the probability that exactly 17 seeds germinate. Show your work.

(b) If only 12 seeds actually germinate, should Judy be suspicious that the company's claim is not true? Compute $P(X \leq 12)$ and use this result to support your answer.

80. **Taking the train** Refer to Exercise 72.

(a) Find the probability that the train arrives late on exactly 2 days. Show your work.

(b) Would you be surprised if the train arrived late on 2 or more days? Compute $P(W \geq 2)$ and use this result to support your answer.

81. **Random digit dialing** When an opinion poll calls a residential telephone number at random, there is only a 20% chance that the call reaches a live person. You watch the random digit dialing machine make 15 calls. Let X = the number of calls that reach a live person.

(a) Find and interpret μ_X.

(b) Find and interpret σ_X.

82. **Lie detectors** A federal report finds that lie detector tests given to truthful persons have probability about 0.2 of suggesting that the person is deceptive.[12] A company asks 12 job applicants about thefts from previous employers, using a lie detector to assess their truthfulness. Suppose that all 12 answer truthfully. Let X = the number of people who the lie detector says are being deceptive.

(a) Find and interpret μ_X.

(b) Find and interpret σ_X.

83. **Random digit dialing** Refer to Exercise 81. Let Y = the number of calls that *don't* reach a live person.

(a) Find the mean of Y. How is it related to the mean of X? Explain why this makes sense.

(b) Find the standard deviation of Y. How is it related to the standard deviation of X? Explain why this makes sense.

84. **Lie detectors** Refer to Exercise 82. Let Y = the number of people who the lie detector says are telling the truth.

(a) Find $P(Y \geq 10)$. How is this related to $P(X \leq 2)$? Explain.

(b) Calculate μ_Y and σ_Y. How do they compare with μ_X and σ_X? Explain why this makes sense.

85. **1 in 6 wins** As a special promotion for its 20-ounce bottles of soda, a soft drink company printed a message on the inside of each cap. Some of the caps said, "Please try again," while others said, "You're a winner!" The company advertised the promotion with the slogan "1 in 6 wins a prize." Suppose the company is telling the truth and that every 20-ounce bottle of soda it fills has a 1-in-6 chance of being a winner. Seven friends each buy one 20-ounce bottle of the soda at a local convenience store. Let X = the number who win a prize.

pg 399

(a) Explain why X is a binomial random variable.

(b) Find the mean and standard deviation of X. Interpret each value in context.

(c) The store clerk is surprised when three of the friends win a prize. Is this group of friends just lucky, or is the company's 1-in-6 claim inaccurate? Compute $P(X \geq 3)$ and use the result to justify your answer.

86. **Aircraft engines** Engineers define reliability as the probability that an item will perform its function under specific conditions for a specific period of time. A certain model of aircraft engine is designed so that each engine has probability 0.999 of performing properly for an hour of flight. Company engineers test an SRS of 350 engines of this model. Let X = the number that operate for an hour without failure.

(a) Explain why X is a binomial random variable.

(b) Find the mean and standard deviation of X. Interpret each value in context.

(c) Two engines failed the test. Are you convinced that this model of engine is less reliable than it's supposed to be? Compute $P(X \leq 348)$ and use the result to justify your answer.

87. **Airport security** The Transportation Security Administration (TSA) is responsible for airport safety. On some flights, TSA officers randomly select passengers for an extra security check before boarding. One such flight had 76 passengers—12 in first class and 64 in coach class. Some passengers were surprised when none of the 10 passengers chosen for screening were seated in first class. Can we use a binomial distribution to approximate this probability? Justify your answer.

pg 402

88. **Scrabble** In the game of Scrabble, each player begins by drawing 7 tiles from a bag containing 100 tiles. There are 42 vowels, 56 consonants, and 2 blank tiles in the bag. Cait chooses her 7 tiles and is surprised to discover that all of them are vowels. Can we use a binomial distribution to approximate this probability? Justify your answer.

89. **10% condition** To use a binomial distribution to approximate the count of successes in an SRS, why do we require that the sample size n be no more than 10% of the population size N?

90. ***Large Counts condition** To use a Normal distribution to approximate binomial probabilities, why do we require that both np and $n(1 - p)$ be at least 10?

91. ***On the Web** What kinds of Web sites do males aged 18 to 34 visit most often? Half of male Internet users in this age group visit an auction site such as eBay at least once a month.[13] A study of Internet use interviews a random sample of 500 men aged 18 to 34. Let X = the number in the sample who visit an auction site at least once a month.

pg 403

(a) Show that X is approximately a binomial random variable.

(b) Check the conditions for using a Normal approximation in this setting.

(c) Use a Normal distribution to estimate the probability that at least 235 of the men in the sample visit an online auction site at least once a month.

92. ***Checking for survey errors** One way of checking the effect of undercoverage, nonresponse, and other sources of error in a sample survey is to compare the sample with known facts about the population. About 12% of American adults identify themselves as black. Suppose we take an SRS of 1500 American adults and let X be the number of blacks in the sample.

(a) Show that X is approximately a binomial random variable.

(b) Check the conditions for using a Normal approximation in this setting.

(c) Use a Normal distribution to estimate the probability that the sample will contain between 165 and 195 blacks.

93. **Using Benford's law** According to Benford's law (Exercise 5, page 359), the probability that the first digit of the amount on a randomly chosen invoice is a 1 or a 2 is 0.477. Suppose you examine an SRS of 90 invoices from a vendor and find 29 that have first digits 1 or 2. Do you suspect that the invoice

*This topic is not required for the AP® Statistics exam. Some teachers prefer to discuss this topic when presenting the sampling distribution of $\hat{p}$ (Chapter 7).

amounts are not genuine? Compute an appropriate probability to support your answer.

94. **A .300 hitter** In baseball, a 0.300 hitter gets a hit in 30% of times at bat. When a baseball player hits 0.300, fans tend to be impressed. Typical Major Leaguers bat about 500 times a season and hit about 0.260. A hitter's successive tries seem to be independent. Could a typical Major Leaguer hit 0.300 just by chance? Compute an appropriate probability to support your answer.

95. **Geometric or not?** Determine whether each of the following scenarios describes a geometric setting. If so, define an appropriate geometric random variable.

(a) A popular brand of cereal puts a card with 1 of 5 famous NASCAR drivers in each box. There is a 1/5 chance that any particular driver's card ends up in any box of cereal. Buy boxes of the cereal until you have all 5 drivers' cards.

(b) In a game of 4-Spot Keno, Lola picks 4 numbers from 1 to 80. The casino randomly selects 20 winning numbers from 1 to 80. Lola wins money if she picks 2 or more of the winning numbers. The probability that this happens is 0.259. Lola decides to keep playing games of 4-Spot Keno until she wins some money.

96. **Geometric or not?** Determine whether each of the following scenarios describes a geometric setting. If so, define an appropriate geometric random variable.

(a) Shuffle a standard deck of playing cards well. Then turn over one card at a time from the top of the deck until you get an ace.

(b) Lawrence likes to shoot a bow and arrow in his free time. On any shot, he has about a 10% chance of hitting the bull's-eye. As a challenge one day, Lawrence decides to keep shooting until he gets a bull's-eye.

97. **1-in-6 wins** Alan decides to use a different strategy for the 1-in-6 wins game of Exercise 85. He keeps buying one 20-ounce bottle of the soda at a time until he gets a winner.

pg 406

(a) Find the probability that he buys exactly 5 bottles. Show your work.

(b) Find the probability that he buys no more than 8 bottles. Show your work.

98. **Cranky mower** To start her old lawn mower, Rita has to pull a cord and hope for some luck. On any particular pull, the mower has a 20% chance of starting.

(a) Find the probability that it takes her exactly 3 pulls to start the mower. Show your work.

(b) Find the probability that it takes her more than 10 pulls to start the mower. Show your work.

99. **Using Benford's law** According to Benford's law (Exercise 5, page 359), the probability that the first digit of the amount of a randomly chosen invoice is an 8 or a 9 is 0.097. Suppose you examine randomly selected invoices from a vendor until you find one whose amount begins with an 8 or a 9.

(a) How many invoices do you expect to examine until you get one that begins with an 8 or 9? Justify your answer.

(b) In fact, you don't get an amount starting with an 8 or 9 until the 40th invoice. Do you suspect that the invoice amounts are not genuine? Compute an appropriate probability to support your answer.

100. **Roulette** Marti decides to keep placing a \$1 bet on number 15 in consecutive spins of a roulette wheel until she wins. On any spin, there's a 1-in-38 chance that the ball will land in the 15 slot.

(a) How many spins do you expect it to take until Marti wins? Justify your answer.

(b) Would you be surprised if Marti won in 3 or fewer spins? Compute an appropriate probability to support your answer.

Multiple choice: Select the best answer for Exercises 101 to 105.

101. Joe reads that 1 out of 4 eggs contains salmonella bacteria. So he never uses more than 3 eggs in cooking. If eggs do or don't contain salmonella independently of each other, the number of contaminated eggs when Joe uses 3 chosen at random has the following distribution:

(a) binomial; $n = 4$ and $p = 1/4$

(b) binomial; $n = 3$ and $p = 1/4$

(c) binomial; $n = 3$ and $p = 1/3$

(d) geometric; $p = 1/4$

(e) geometric; $p = 1/3$

Exercises 102 and 103 refer to the following setting. A fast-food restaurant runs a promotion in which certain food items come with game pieces. According to the restaurant, 1 in 4 game pieces is a winner.

102. If Jeff gets 4 game pieces, what is the probability that he wins exactly 1 prize?

(a) 0.25

(b) 1.00

(c) $\binom{4}{1}(0.25)^1(0.75)^3$

(d) $\binom{4}{1}(0.25)^3(0.75)^1$

(e) $(0.75)^3(0.25)^1$

103. If Jeff keeps playing until he wins a prize, what is the probability that he has to play the game exactly 5 times?

(a) $(0.25)^5$
(b) $(0.75)^4$
(c) $(0.75)^5$
(d) $(0.75)^4(0.25)$
(e) $\binom{5}{1}(0.75)^4(0.25)$

104. Each entry in a table of random digits like Table D has probability 0.1 of being a 0, and the digits are independent of one another. If many lines of 40 random digits are selected, the mean and standard deviation of the number of 0s will be approximately

(a) mean = 0.1, standard deviation = 0.05.
(b) mean = 0.1, standard deviation = 0.1.
(c) mean = 4, standard deviation = 0.05.
(d) mean = 4, standard deviation = 1.90.
(e) mean = 4, standard deviation = 3.60.

105. *In which of the following situations would it be appropriate to use a Normal distribution to approximate probabilities for a binomial distribution with the given values of n and p?

(a) $n = 10, p = 0.5$
(b) $n = 40, p = 0.88$
(c) $n = 100, p = 0.2$
(d) $n = 100, p = 0.99$
(e) $n = 1000, p = 0.003$

*This topic is not required for the AP® Statistics exam. Some teachers prefer to discuss this topic when presenting the sampling distribution of $\hat{p}$ (Chapter 7).

106. Spoofing (4.2) To collect information such as passwords, online criminals use "spoofing" to direct Internet users to fraudulent Web sites. In one study of Internet fraud, students were warned about spoofing and then asked to log in to their university account starting from the university's home page. In some cases, the login link led to the genuine dialog box. In others, the box looked genuine but in fact was linked to a different site that recorded the ID and password the student entered. The box that appeared for each student was determined at random. An alert student could detect the fraud by looking at the true Internet address displayed in the browser status bar, but most just entered their ID and password. Is this study an experiment? Why? What are the explanatory and response variables?

107. Smoking and social class (5.3) As the dangers of smoking have become more widely known, clear class differences in smoking have emerged. British government statistics classify adult men by occupation as "managerial and professional" (43% of the population), "intermediate" (34%), or "routine and manual" (23%). A survey finds that 20% of men in managerial and professional occupations smoke, 29% of the intermediate group smoke, and 38% in routine and manual occupations smoke.[14]

(a) Use a tree diagram to find the percent of all adult British men who smoke.
(b) Find the percent of male smokers who have routine and manual occupations.

FRAPPY! Free Response AP® Problem, Yay!

The following problem is modeled after actual AP® Statistics exam free response questions. Your task is to generate a complete, concise response in 15 minutes.

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

Buckley Farms produces homemade potato chips that it sells in bags labeled "16 ounces." The total weight of each bag follows an approximately Normal distribution with a mean of 16.15 ounces and a standard deviation of 0.12 ounces.

(a) If you randomly selected 1 bag of these chips, what is the probability that the total weight is less than 16 ounces?
(b) If you randomly selected 10 bags of these chips, what is the probability that exactly 2 of the bags will have a total weight less than 16 ounces?
(c) Buckley Farms ships its chips in boxes that contain 6 bags. The empty boxes have a mean weight of 10 ounces and a standard deviation of 0.05 ounces. Calculate the mean and standard deviation of the total weight of a box containing 6 bags of chips.
(d) Buckley Farms decides to increase the mean weight of each bag of chips so that only 5% of the bags have weights that are less than 16 ounces. Assuming that the standard deviation remains 0.12 ounces, what mean weight should Buckley Farms use?

After you finish, you can view two example solutions on the book's Web site (www.whfreeman.com/tps5e). Determine whether you think each solution is "complete," "substantial," "developing," or "minimal." If the solution is not complete, what improvements would you suggest to the student who wrote it? Finally, your teacher will provide you with a scoring rubric. Score your response and note what, if anything, you would do differently to improve your own score.

Chapter Review

Section 6.1: Discrete and Continuous Random Variables

A random variable assigns numerical values to the outcomes of a chance process. The probability distribution of a random variable describes its possible values and their probabilities. There are two types of random variables: discrete and continuous. Discrete random variables take on a fixed set of values with gaps in between. Continuous random variables take on all values in an interval of numbers.

As in Chapter 1, we are often interested in the shape, center, and spread of a probability distribution. The shape of a discrete probability distribution can be identified by graphing a probability histogram, with the height of each bar representing the probability of a single value. The center is usually identified by the mean (expected value) of the random variable. The expected value is the average value of the random variable if the chance process is repeated many times. The spread of a probability distribution is usually identified by the standard deviation, which describes how much the values of a random variable typically differ from the mean value, in many repetitions of the chance process.

Continuous probability distributions, such as the Normal distribution, describe the distribution of continuous random variables. A density curve is used to display a continuous probability distribution. Probabilities for continuous random variables are determined by finding the area under the density curve and above the values of interest.

Section 6.2: Transforming and Combining Random Variables

In this section, you learned how linear transformations of a random variable affect the shape, center, and spread of its probability distribution. As you learned in Chapter 2, a linear transformation does not change the shape (unless you multiply by a negative number) but can change the center and spread depending on the type of transformation. Multiplying (or dividing) each value of the random variable by a positive constant b multiplies (divides) the mean and standard deviation by b. Adding a constant a to (subtracting a from) each value of the random variable adds a to (subtracts a from) the mean but doesn't change the standard deviation.

You also learned how to calculate the mean and standard deviation for a combination of two or more random variables. If you are adding two random variables, X and Y, the mean and standard deviation of X + Y are

$$\mu_{X+Y} = \mu_X + \mu_Y \quad \text{and} \quad \sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$$

Likewise, if you are subtracting two random variables, X and Y, the mean and standard deviation of X − Y are

$$\mu_{X-Y} = \mu_X - \mu_Y \quad \text{and} \quad \sigma_{X-Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$$

The formulas for the standard deviation of X + Y and X − Y are only correct if X and Y are independent, that is, if knowing the value of one variable doesn't provide any additional information about the other variable. Also, if X and Y are both Normally distributed, then X + Y and X − Y are both Normally distributed as well.

To determine which formulas to use for a particular problem, it is important to be able to distinguish linear transformations and combinations of random variables. Linear transformations take the values of *one* random variable and add, subtract, multiply, or divide them by a constant. Combinations of random variables take *two or more* random variables and add or subtract them. When a problem involves both linear transformations and a combination of random variables, remember to do the linear transformations first.

Section 6.3: Binomial and Geometric Random Variables

In this section, you learned about two common types of discrete random variables, binomial random variables and geometric random variables. Binomial random variables count the number of successes in a fixed number of trials (n), whereas geometric random variables count the number of trials needed to get one success. Otherwise, the binomial and geometric settings have the same conditions: there must be two possible outcomes for each trial (success or failure), the trials must be independent, and the probability of success p must stay the same throughout all trials.

To calculate probabilities for a binomial distribution with n trials and probability of success p on each trial, use technology or the binomial probability formula

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

The mean and standard deviation of a binomial random variable X are

$$\mu_X = np \quad \text{and} \quad \sigma_X = \sqrt{np(1 - p)}$$

The shape of a binomial distribution depends on both the number of trials n and the probability of success p. When the number of trials is large enough that both np and $n(1-p)$ are at least 10, the distribution of the binomial random variable X has an approximately Normal distribution. Be sure to check the large counts condition before using a Normal approximation to a binomial distribution.

A common application of the binomial distribution is when we count the number of times a particular outcome occurs in a random sample from some population. Because sampling is almost always done without replacement, the independence condition is violated. However, if the sample size is a small fraction of the population size (less than 10%), the lack of independence isn't a concern. Be sure to check the 10% condition when sampling is done without replacement before using a binomial distribution.

Finally, to calculate probabilities for a geometric distribution with probability of success p on each trial, use technology or the geometric probability formula

$$P(Y = k) = (1 - p)^{k-1} p$$

What Did You Learn?

Learning Objective	Section	Related Example on Page(s)	Relevant Chapter Review Exercise(s)
Compute probabilities using the probability distribution of a discrete random variable.	6.1	349	R6.1
Calculate and interpret the mean (expected value) of a discrete random variable.	6.1	350, 352	R6.1, R6.3
Calculate and interpret the standard deviation of a discrete random variable.	6.1	353	R6.1, R6.3
Compute probabilities using the probability distribution of certain continuous random variables.	6.1	355, 357	R6.4
Describe the effects of transforming a random variable by adding or subtracting a constant and multiplying or dividing by a constant.	6.2	365, 366, 368	R6.2, R6.3
Find the mean and standard deviation of the sum or difference of independent random variables.	6.2	372, 373, 374, 377	R6.3, R6.4
Find probabilities involving the sum or difference of independent Normal random variables.	6.2	380, 381	R6.4
Determine whether the conditions for using a binomial random variable are met.	6.3	388	R6.5
Compute and interpret probabilities involving binomial distributions.	6.3	390, 393, 396	R6.6
Calculate the mean and standard deviation of a binomial random variable. Interpret these values in context.	6.3	399	R6.5
Find probabilities involving geometric random variables.	6.3	406	R6.7
*When appropriate, use the Normal approximation to the binomial distribution to calculate probabilities.	6.3	403	R6.8

*This topic is not required for the AP® Statistics exam.

Chapter 6 Chapter Review Exercises

These exercises are designed to help you review the important ideas and methods of the chapter.

R6.1 Knees Patients receiving artificial knees often experience pain after surgery. The pain is measured on a subjective scale with possible values of 1 (low) to 5 (high). Let X be the pain score for a randomly selected patient. The following table gives part of the probability distribution for X.

Value:	1	2	3	4	5
Probability:	0.1	0.2	0.3	0.3	??

(a) Find $P(X = 5)$.

(b) Is pain score a discrete or continuous random variable? Explain.

(c) Find $P(X \leq 2)$. Is this the same as $P(X < 2)$? Explain.

(d) Compute the expected pain score and the standard deviation of the pain scores.

R6.2 A glass act In a process for manufacturing glassware, glass stems are sealed by heating them in a flame. Let X be the temperature (in degrees Celsius) for a randomly chosen glass. The mean and standard deviation of X are $\mu_X = 550°\text{C}$ and $\sigma_X = 5.7°\text{C}$.

(a) Is temperature a discrete or continuous random variable? Explain.
(b) How is $P(X < 540)$ related to $P(X \leq 540)$? Explain.
(c) The target temperature is 550°C. What are the mean and standard deviation of the number of degrees off target, $D = X - 550$?
(d) A manager asks for results in degrees Fahrenheit. The conversion of X into degrees Fahrenheit is given by $Y = \frac{9}{5}X + 32$. What are the mean μ_Y and the standard deviation σ_Y of the temperature of the flame in the Fahrenheit scale?

R6.3 Keno In a game of 4-Spot Keno, the player picks 4 numbers from 1 to 80. The casino randomly selects 20 winning numbers from 1 to 80. The table below shows the possible outcomes of the game and their probabilities, along with the amount of money (Payout) that the player wins for a \$1 bet. If X = the payout for a single \$1 bet, you can check that $\mu_X =$ \$0.70 and $\sigma_X =$ \$6.58.

Matches:	0	1	2	3	4
Payout x_i:	\$0	\$0	\$1	\$3	\$120
Probability p_i:	0.308	0.433	0.213	0.043	0.003

(a) Interpret the values of μ_X and σ_X in context.
(b) Jerry places a single \$5 bet on 4-Spot Keno. Find the expected value and the standard deviation of his winnings.
(c) Marla plays five games of 4-Spot Keno, betting \$1 each time. Find the expected value and the standard deviation of her total winnings.
(d) Based on your answers to (b) and (c), which player would the casino prefer? Justify your answer.

R6.4 Applying torque A machine fastens plastic screw-on caps onto containers of motor oil. If the machine applies more torque than the cap can withstand, the cap will break. Both the torque applied and the strength of the caps vary. The capping-machine torque T follows a Normal distribution with mean 7 inch-pounds and standard deviation 0.9 inch-pounds. The cap strength C (the torque that would break the cap) follows a Normal distribution with mean 10 inch-pounds and standard deviation 1.2 inch-pounds.

(a) What is the probability that a randomly selected cap has a strength greater than 11 inch-pounds?
(b) Explain why it is reasonable to assume that the cap strength and the torque applied by the machine are independent.
(c) Let the random variable $D = C - T$. Find its mean and standard deviation.
(d) What is the probability that a randomly selected cap will break while being fastened by the machine? Show your work.

Exercises R6.5 and R6.6 refer to the following setting. According to Mars, Incorporated, 20% of its plain M&M'S candies are orange. Assume that the company's claim is true. Suppose that you reach into a large bag of plain M&M'S (without looking) and pull out 8 candies. Let X = the number of orange candies you get.

R6.5 Orange M&M'S

(a) Explain why it is reasonable to use the binomial distribution for probability calculations involving X.
(b) Find and interpret the expected value of X.
(c) Find and interpret the standard deviation of X.

R6.6 Orange M&M'S

(a) Would you be surprised if none of the candies were orange? Compute an appropriate probability to support your answer.
(b) How surprising would it be to get 5 or more orange candies? Compute an appropriate probability to support your answer.

R6.7 *Sushi Roulette* In the Japanese game show *Sushi Roulette*, the contestant spins a large wheel that's divided into 12 equal sections. Nine of the sections have a sushi roll, and three have a "wasabi bomb." When the wheel stops, the contestant must eat whatever food is on that section. To win the game, the contestant must eat one wasabi bomb. Find the probability that it takes 3 or fewer spins for the contestant to get a wasabi bomb. Show your method clearly.

R6.8* Is this coin balanced? While he was a prisoner of war during World War II, John Kerrich tossed a coin 10,000 times. He got 5067 heads. If the coin is perfectly balanced, the probability of a head is 0.5.

(a) Find the mean and the standard deviation of the number of heads in 10,000 tosses, assuming the coin is perfectly balanced.
(b) Explain why the Normal approximation is appropriate for calculating probabilities involving the number of heads in 10,000 tosses.
(c) Is there reason to think that Kerrich's coin was not balanced? To answer this question, use a Normal distribution to estimate the probability that tossing a balanced coin 10,000 times would give a count of heads at least this far from 5000 (that is, at least 5067 heads or at most 4933 heads).

*This topic is not required for the AP® Statistics exam. Some teachers prefer to discuss this topic when presenting the sampling distribution of $\hat{p}$ (Chapter 7).

Chapter 6 AP® Statistics Practice Test

Section I: Multiple Choice *Select the best answer for each question.*

Questions T6.1 to T6.3 refer to the following setting. A psychologist studied the number of puzzles that subjects were able to solve in a five-minute period while listening to soothing music. Let X be the number of puzzles completed successfully by a randomly chosen subject. The psychologist found that X had the following probability distribution:

Value:	1	2	3	4
Probability:	0.2	0.4	0.3	0.1

T6.1 What is the probability that a randomly chosen subject completes more than the expected number of puzzles in the five-minute period while listening to soothing music?

(a) 0.1
(b) 0.4
(c) 0.8
(d) 1
(e) Cannot be determined

T6.2 The standard deviation of X is 0.9. Which of the following is the best interpretation of this value?

(a) About 90% of subjects solved 3 or fewer puzzles.
(b) About 68% of subjects solved between 0.9 puzzles less and 0.9 puzzles more than the mean.
(c) The typical subject solved an average of 0.9 puzzles.
(d) The number of puzzles solved by subjects typically differed from the mean by about 0.9 puzzles.
(e) The number of puzzles solved by subjects typically differed from one another by about 0.9 puzzles.

T6.3 Let D be the difference in the number of puzzles solved by two randomly selected subjects in a five-minute period. What is the standard deviation of D?

(a) 0 (b) 0.81 (c) 0.9 (d) 1.27 (e) 1.8

T6.4 Suppose a student is randomly selected from your school. Which of the following pairs of random variables are most likely independent?

(a) X = student's height; Y = student's weight
(b) X = student's IQ; Y = student's GPA
(c) X = student's PSAT Math score; Y = student's PSAT Verbal score
(d) X = average amount of homework the student does per night; Y = student's GPA
(e) X = average amount of homework the student does per night; Y = student's height

T6.5 A certain vending machine offers 20-ounce bottles of soda for \$1.50. The number of bottles X bought from the machine on any day is a random variable with mean 50 and standard deviation 15. Let the random variable Y equal the total revenue from this machine on a given day. Assume that the machine works properly and that no sodas are stolen from the machine. What are the mean and standard deviation of Y?

(a) $\mu_Y = \$1.50, \sigma_Y = \$22.50$
(b) $\mu_Y = \$1.50, \sigma_Y = \$33.75$
(c) $\mu_Y = \$75, \sigma_Y = \$18.37$
(d) $\mu_Y = \$75, \sigma_Y = \$22.50$
(e) $\mu_Y = \$75, \sigma_Y = \$33.75$

T6.6 The weight of tomatoes chosen at random from a bin at the farmer's market follows a Normal distribution with mean $\mu = 10$ ounces and standard deviation $\sigma = 1$ ounce. Suppose we pick four tomatoes at random from the bin and find their total weight T. The random variable T is

(a) Normal, with mean 10 ounces and standard deviation 1 ounce.
(b) Normal, with mean 40 ounces and standard deviation 2 ounces.
(c) Normal, with mean 40 ounces and standard deviation 4 ounces.
(d) binomial, with mean 40 ounces and standard deviation 2 ounces.
(e) binomial, with mean 40 ounces and standard deviation 4 ounces.

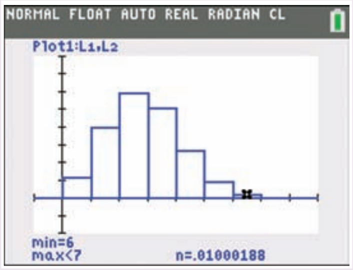
T6.7 Which of the following random variables is geometric?

(a) The number of times I have to roll a die to get two 6s.
(b) The number of cards I deal from a well-shuffled deck of 52 cards until I get a heart.
(c) The number of digits I read in a randomly selected row of the random digits table until I find a 7.
(d) The number of 7s in a row of 40 random digits.
(e) The number of 6s I get if I roll a die 10 times.

T6.8 Seventeen people have been exposed to a particular disease. Each one independently has a 40% chance of contracting the disease. A hospital has the capacity to handle 10 cases of the disease. What is the probability that the hospital's capacity will be exceeded?

(a) 0.011
(b) 0.035
(c) 0.092
(d) 0.965
(e) 0.989

T6.9 The figure shows the probability distribution of a discrete random variable X. Note that the cursor is on the histogram bar representing a value of 6. Which of the following best describes this random variable?



(a) Binomial with $n = 8, p = 0.1$
(b) Binomial with $n = 8, p = 0.3$
(c) Binomial with $n = 8, p = 0.8$
(d) Geometric with $p = 0.1$
(e) Geometric with $p = 0.2$

T6.10 A test for extrasensory perception (ESP) involves asking a person to tell which of 5 shapes—a circle, star, triangle, diamond, or heart—appears on a hidden computer screen. On each trial, the computer is equally likely to select any of the 5 shapes. Suppose researchers are testing a person who does not have ESP and so is just guessing on each trial. What is the probability that the person guesses the first 4 shapes incorrectly but gets the fifth correct?

(a) 1/5
(b) $\left(\frac{4}{5}\right)^4$
(c) $\left(\frac{4}{5}\right)^4 \cdot \left(\frac{1}{5}\right)$
(d) $\binom{5}{1} \cdot \left(\frac{4}{5}\right)^4 \cdot \left(\frac{1}{5}\right)$
(e) 4/5

Section II: Free Response *Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.*

T6.11 Let Y denote the number of broken eggs in a randomly selected carton of one dozen "store brand" eggs at a local supermarket. Suppose that the probability distribution of Y is as follows.

Value y_i:	0	1	2	3	4
Probability p_i:	0.78	0.11	0.07	0.03	0.01

(a) What is the probability that at least 10 eggs in a randomly selected carton are *unbroken*?
(b) Calculate and interpret μ_Y.
(c) Calculate and interpret σ_Y. Show your work.
(d) A quality control inspector at the store keeps looking at randomly selected cartons of eggs until he finds one with at least 2 broken eggs. Find the probability that this happens in one of the first three cartons he inspects.

T6.12 *Ladies Home Journal* magazine reported that 66% of all dog owners greet their dog before greeting their spouse or children when they return home at the end of the workday. Assume that this claim is true. Suppose 12 dog owners are selected at random. Let X = the number of owners who greet their dogs first.

(a) Explain why it is reasonable to use the binomial distribution for probability calculations involving X.
(b) Only 4 of the owners in the random sample greeted their dogs first. Does this give convincing evidence against the *Ladies Home Journal* claim? Calculate an appropriate probability to support your answer.

T6.13 Ed and Adelaide attend the same high school, but are in different math classes. The time E that it takes Ed to do his math homework follows a Normal distribution with mean 25 minutes and standard deviation 5 minutes. Adelaide's math homework time A follows a Normal distribution with mean 50 minutes and standard deviation 10 minutes. Assume that E and A are independent random variables.

(a) Randomly select one math assignment of Ed's and one math assignment of Adelaide's. Let the random variable D be the difference in the amount of time each student spent on their assignments: $D = A - E$. Find the mean and the standard deviation of D. Show your work.
(b) Find the probability that Ed spent longer on his assignment than Adelaide did on hers. Show your work.

T6.14 According to the Census Bureau, 13% of American adults (aged 18 and over) are Hispanic. An opinion poll plans to contact an SRS of 1200 adults.

(a) What is the mean number of Hispanics in such samples? What is the standard deviation?
(b) Should we be suspicious if the sample selected for the opinion poll contains 15% Hispanic people? Compute an appropriate probability to support your answer.